Finding the Road to Safety: A Predictive Approach to Traffic Fatalities in the United States

Joshua Rosen
McCourt School of Public Policy, Georgetown University

December 16, 2021

Abstract

This paper attempts to produce a high-level description of the variables correlated with traffic fatalities in the United States. Ultimately, a predictive modeling approach is utilized to examine the National Highway Traffic Safety Administration's Fatality Analysis Reporting System. The investigation ends with a brief discussion on next-steps to produce policy-relevant findings.

Keywords: Logistic regression, machine learning, urban policy

Word Count: 3000

1 Introduction and Background

Between 2010 and 2020, over 350,000 individuals have been killed in traffic accidents. As a result, US Transportation Secretary Pete Buttigieg has labeled the rapid rise in unnecessary traffic deaths a "crisis," and has called for DOT to produce the first ever National Roadway Safety Strategy to identify discrete steps for minimizing road dangers.[1] Similarly, over the same ten year span during which American's witnessed a 10-percent spike in traffic-related deaths, the European Union experienced a 36-percent drop.[2] A 2019 paper indirectly examined this trend by investigating the causality of traffic fatalities in both high and middle income countries across three continents. In the paper, the author's found traffic fatalities increased by nearly one-percent for every single percent rise in the number of vehicles, and decreased by half-a-percent for each one-percent increase in urbanization.[3] It is therefore evident that America's traffic fatality crisis is largely a function of America's distinct culture, laws, institutions, and infrastructure.

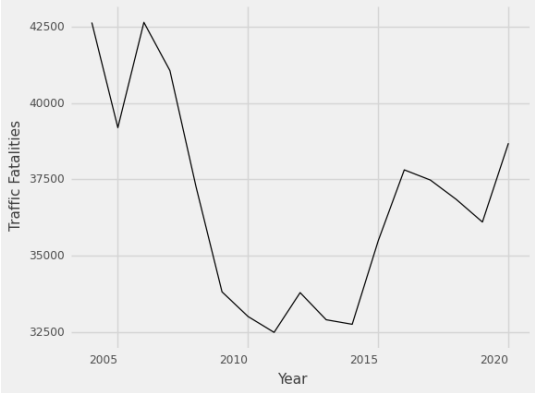

Figure 1: Traffic Fatalities: 2004 to 2020

[1] "USDOT Releases New Data Showing That Road Fatalities Spiked in First Half of 2021," U.S. Department of Transportation, published October 28, 2021, https://www.transportation.gov/briefing-room/usdot-releases-new-data-showing-road-fatalities-spiked-first-half-2021.

[2] "A Grammar of Graphics for Python," plotnine, accessed November 22, 2021, https://plotnine.readthedocs.io/en/stable/.

[3] Ali, Q., Yaseen, M.R. & Khan, M.T.I. "Road traffic fatalities and its determinants in high-income countries: a continent-wise comparison." Environ Sci Pollut Res 26, (2019). https://doi.org/10.1007/s11356-019-05410-9

In addition, in a recent article published in *The Atlantic*, David Zipper, a Visiting Fellow at the Harvard Kennedy School's Taubman Center, criticized the National Highway Traffic Safety Administration's 2015 declaration that 94-percent of crashes are the result of driver failure and human error. Zipper questioned the rationale behind identifying one single determinant of traffic fatalities, rather than properly crediting the complex interaction of systemic practices, institutions, and related variables.[4]

In light of the evolving crisis, the goal of this paper is to formulate a high-level understanding of traffic fatalities in the American context. This broadly defined goal is operationalized into three primary components. First, the project aggregates widely available traffic fatality data into a single, unified source. Second, visualizations are utilized to present visually informative depictions of the core problem.[5] Finally, this paper relies on predictive modeling to pin-point which variables are highly correlated with deadly crashes. Ultimately, despite the construction of a highly predictive logistic regression model, this paper does not attempt to formulate a novel hypothesis/outlook. The paper concludes with a discussion on 'next steps' to better achieve a policy-relevant understanding of traffic fatalities in the United States, as well as possible future avenues for additional research questions.

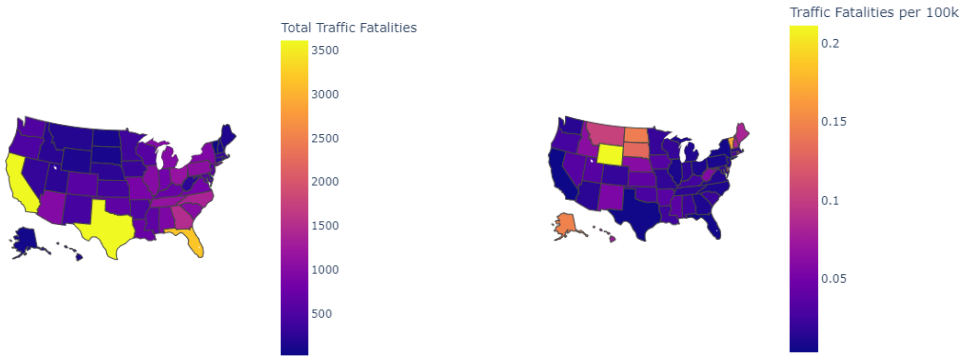

Figure 2: 2019 Traffic Fatalities by State

[4] David Zipper, "The DEadly Myth That Human Error Causes Most Car Crashes," *The Atlantic*, November 26, 2021.

[5] "Dash Enterprise," plotly, accessed November 17, 2021, `https://plotly.com/`.

2 Data

This paper utilizes the National Highway Traffic Safety Administration's (NHTSA) Fatality Analysis Reporting System (FARS) to underpin its core conclusions.[6] As the name suggests, FARS represents a nationwide, public repository for traffic fatalities occurring between 1975 and 2019. Each yearly file contains 19 distinct sub-files categorizing different facets of a traffic fatality.[7] Further, each respective sub-file utilizes a distinct unit-of-analysis. Analysis is conducted on a single, merged CSV that has been constructed to contain the three most relevant sub-files from the 2019 records: the Accident Data File, the Vehicle Data File, and the Person Data File. All files were downloaded as CSVs and read into python using Pandas.[8]

Merging the files was simple, but required an understanding of each data file's respective unit-of-analysis. Starting with the Accident file, each new observation represents a unique crash event, with the variable ST_CASE holding a unique crash identifier for each record. Next, the Vehicle data file adds records for each vehicle involved in a crash, as well as pre-crash classification variables. Each observation in the file therefore corresponds to a unique individual involved in the accident. The Accident and Vehicle files were therefore merged along the ST_CASE variable. Finally, the Person file contains variables for all pedestrians and drivers. Each observation thus corresponds to a unique individual, with only some resulting in a fatality. ST_CASE, the unique crash identifier, was utilized to merge the Person file with the Accident data file. Then, the ST_CASE and VEH_NO (a unique vehicle identifier) were utilized to merge the Person and Vehicle data files.

Notably, this process created overlapping ST_CASE and VEH_No observations, with the same crash and vehicle identifiers present in multiple rows corresponding to each person and vehicle involved in the accident. As a result, a simple summation of the fatality variable would result in a massive over-counting of total 2019 fatalities. Further, as the repositories name suggests, the system does not track total traffic *accidents*, but only events

[6] "Fatality Analysis Reporting System," NHTSA, accessed October 27, 2021, `https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars`.

[7] "Fatality Analysis Reporting System (FARS) Analytical User's Manual, 1975-2019," NHTSA, revised February 2021, `https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813023`.

[8] "Pandas," Pandas, accessed December 9, 2021, `https://pandas.pydata.org/`.

that end in at least one *fatality*. In other words, each observation pulled from the Accident data file contains a minimum of one fatality. This facet presented two challenges. First, the visualization component would be severely limited, with no ability to compare the relationship between total accidents and total fatalities. Second, the classification model could not be used to simply categorize which automobile accidents end in a fatality versus which accidents merely end in a few scratches.

Once unified, the data revealed hundreds of possible predictor variables on factors such as: geography, time/date, the driver's previous accident(s), driver demographic variables, weather/light conditions, lane/system types, pre-crash critical events, vehicle make/model, and an individual's location within the automobile. To more efficiently select useful columns, an SQLite database was created and queried through python with variables then manually selected.[9] After transforming the array back to a Pandas Dataframe, this initial clean-through was saved as a new CSV representing the official unified source.

In total, 88 predictor variables were isolated. Since the vast majority of variables were categorical, these fields were converted to dummy variables, with each possible outcome transformed into a single column holding a binary, 'yes' or 'no' distinction. For example, the original RUR_URBNAME variable held possible outcomes of "Rural," "Urban," "Unknown," and "Not Reported." However, after conversion to a dichotomous form, the variable was replaced by four columns indicating either a zero or a one. This process was also utilized to convert time/date values (hour, day, month), thus eliminating the need for time-series' capable models. Altogether, the unified dataframe contained 82220 rows and 2814 columns.

Next, the outcome variable was selected. Since the data does not supply the total population of traffic accidents in a given year, the analysis instead shifted to identify the combination of events that strongly correlate with a "fatal" injury result. In other words, since the data contains each person involved in a given accident, the model aimed to predict which individual would be killed. The injury-type variable produces five possible outcomes: (1) died prior to the crash, (2) fatal injury, (3) injury, severity unknown, (4) no apparent injury, (5) possible injury. As with the predictor variables, the outcome variable

[9] "What Is SQLite?" SQLite, accessed November 10, 2021, `https://www.sqlite.org/index.html`

was similarly transformed to a dummy variable to account for categorical values. Further, all but the fatal injury type outcome were excluded. Figure X displays the distribution of injury outcomes in the data prior to removing all values aside from fatalities.

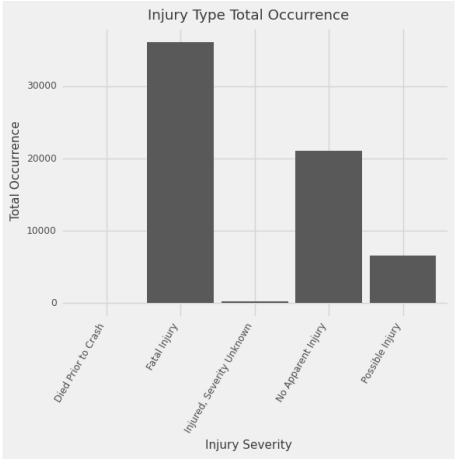

Figure 3: Distribution of Injury Severity

3 Analysis

Analysis was conducted through two methodological approaches. First, visualizations are utilized to presents visually informative depictions of the core problems. However, as will be discussed in later sections, these plots are severely limited by gaps in the original data. Thus, predictive modeling is utilized as the core analytical framework.

Second, supervised statistical learning is undertaken to describe the interactions strongly correlated with traffic fatalities. In more specific terms, statistical learning models aimed to predict which individual(s) involved in the accident were likely to be killed. Other tools were then utilized to select on which particular variables the model leaned to produce its prediction—thus lending insight into the variables most highly correlated with a desired outcome. Since the outcome variable assigned an observation to two possible binary values of no natural ordering—a fatality or not a fatality—the data necessitated using classification models rather than tools better suited for continuous observations. Five classification models were thus utilized at various tuning parameters: (1) Naive Bayes, (2) K-Nearest Neighbors (KNN), (3) Decision Tree, (4) Random Forest, (5) Logistic Regression. An explanation for each model is included below.

First, a Bayes classifier—named for the titular formula on which it is based—allows us to calculated a binary outcome given predictor variables. In other words, the model finds: $P(InjuryLevelFatality = 1 \mid X_i)$. The *Naive* Bayes Classifier used here imposes a simple assumption: that each variable is independent. From this, the model can estimate:

$$P(InjuryLevelFatality = 1 \mid X_i) = P(X_1 \mid InjuryLevelFatality = 1) \times P(X_2 \mid InjuryLevelFatality = 1) \times ... \times P(InjuryLevelFatality = 1)$$

Second, KNN forms its classification based on which points are *nearest* to a set of K *neighbors*. Here, distance is measured in Euclidean distance and K represents a tuning parameter with no correct answer. Tuning parameters are discussed later in this section. As an example, if K is set to 4, then the model bases its classification on the 4 nearest neighbors.

Third, classification trees utilize recursive binary splitting to perform a series of consecutive selections until the desired outcome is reached. The length of the tree—the number of decisions—represents a tuning parameter. Short trees with fewer splits may under-fit, while deeper trees using many splits could over-fit. Decision Trees construct just one tree, while Random Forests grow multiple trees and average their predictive outcome.

Finally, the strength of logistic regression is it's ability to capture latent variables, thereby reflecting an unobserved continuous variable detailing the propensity of an individual outcome observation to equal one. The resulting effect, formed by an *S-shaped* curve, ensures fitted predictions always lie between 0 and 1. Predictions are then conducted through computing $P(InjuryLevelFatality = 1)$ for each given predictor variable and accompanying level.[10]

$$P(InjuryLevelFatality = 1) = \frac{e^{\beta_0 + \beta_1 X_{1i} + ... + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_{1i} + ... + \beta_n X_n}}$$

Each model thus provides a worthwhile method for classification. Scikit-learn was utilized to run each model—thereby allowing for the selection of the model with the highest predictive performance. Similarly, GridSearch and Cross-validation were utilized to

[10]James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani. *An Introduction to Statistucal Learning: with Applications in R* (New York: Springer, 2013), 136.

consider multiple tuning parameters and form a selection accordingly. For all tuning parameters considered, please see the project's public repository. Prior to running the data through the pipeline, it was randomly divided into separate dataframes for training and test respectively. Only the training data was analyzed. In addition, both the test and training data were appropriately scaled. For this, continuous data was transformed to mean 0 and variance 1. Scaling thus addresses large disparities between the variables' magnitudes and units and prevents skewed results.

Finally, note that neither methodological approaches, visualizations nor predictive modeling, perform causal inference. Instead, visualizations perform simple aggregations, while the predictive models investigate which variables are highly correlated with traffic fatalities.

4 Model Results

The primary method utilized to indicate model accuracy was the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. The ROC curve plots a comparison of two facets: the true positive rate (selectivity)—which tracks the fraction of fatalities correctly identified, and the false positive rate—which provides the fraction of non-fatalities incorrectly identified. AUC, true to its title, is then used to calculate the total area under the curve.[11]

Based on AUC, the logistic regression model achieved the highest predictive score—.9472—and was thus utilized for further analysis. As seen in Figure 4, the model predicts far above the 50-percent prediction threshold visualized by the dotted line through the plot.

[11] Ibid., 150-152

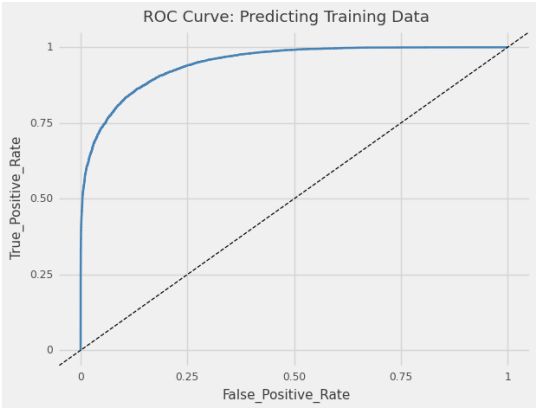

Figure 4: ROC: Training Data

Beyond predictive accuracy, it was essential to interpret (a) on which variables the model leaned most heavily to produce its prediction, and (b) in what direction did individual variables influence the prediction. Three tools were then utilized to achieve the aforementioned goals: (1) permutation feature importance, (2) partial dependence plots (with and without interaction), and (3) global surrogate models.

Permutation feature importance measures the increase/decrease in the prediction's accuracy resulting from a permutation in the prediction variable. Variables with a high, positive *vi* coefficients are thus important for the model's predictive capacity. For example, the severe damage level of the damage extent variable produced a variable importance coefficient of .016006. Given an AUC score of .9472, excluding the severe damage variable level would lower prediction to roughly .931194. Roughly 800 of the nearly 3000 predictor variables reported positive importance coefficients. For a full list of variable importance, please visit the project's public repository. However, permutation feature importance fails to describe in which direction a given variable influences the prediction.

Partial dependency plots and global surrogate models therefore provide necessary insight into model prediction. Starting with the former, partial dependency plots illustrate the marginal effect of one or two variables on the predicted outcome. Figure 5 displays five noteworthy findings derived from this method. The first plot demonstrates how $P(InjuryLevelFatality = 1)$ changes based on the position of an individual in an automobile. Notably, the plots use dummy independent variables—thus reflecting how the prediction changes based on the independent variable's activation. Here, the plot displays that the model is less likely to predict a fatality result for the individual in the driver's

seat. The second plot displays the effect of urban and rural environments on fatality prediction—illustrating that the model is less likely to predict a fatality outcome in an urban environment than a rural environment. While the first two results display a minor, negative relationship, the third finding produces a much stronger, positive prediction when activated. Specifically, the model finds that a vehicle that undergoes severe, disabling damage is strongly correlated with the occurrence of a traffic fatality. Next, the plot indicates that BODY_TYP_x_80.0's involvement (the data's method for classifying motorcycles), aids the model in classifying a fatality. Finally, the model finds that interstates are slightly less correlated with deadly crashes than comparative systems. While the findings require future analysis, this insight nevertheless adds to the data-derived picture of traffic fatalities in the United States.

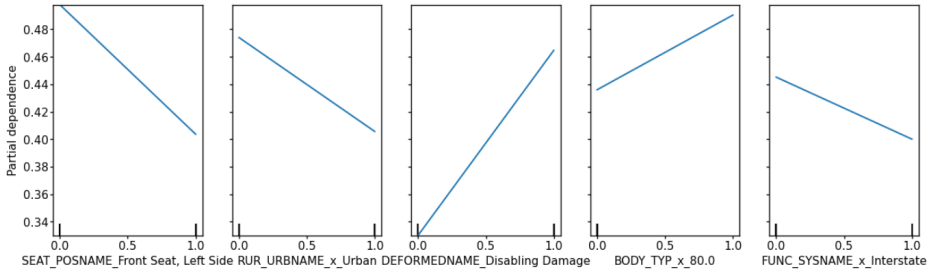

Figure 5: Partial Dependency Plots

Critically, these plots fail to capture interaction between two variables. Below, one notable effect of interaction is thus displayed. Figure 6 displays the interaction between urban environments and pedestrian fatalities. Recall that the RUR_URBNAME_x_Urban variable is used to designate urban environments. Further, PER_TYPENAME_Pedestrian is utilized to denote when the individual killed is a pedestrian. Here, brighter colors distinguish when the model is more likely to predict a fatality. As a result, the plot demonstrates that the model is more likely to predict a pedestrian being killed in rural areas.

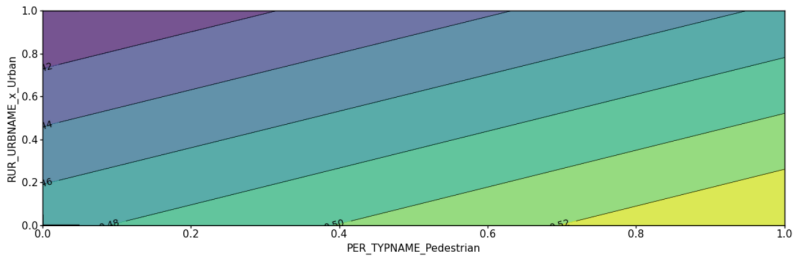

Figure 6: Interaction Between Pedestrian Fatalities and Urban Environments

The final tool utilized for model interpretation is a global surrogate model. In simple terms, a global surrogate model takes the previously generated predictive probabilities for each observation and builds a second model on top. This second model then provides an interpretation via an approximation of the true model's prediction. Here the surrogate model takes the form of a decision tree regressor due to its easily interpretable features. Of note, however, the global surrogate model only achieves an R^2 of .62 and thus fails to produce highly accurate results. In other words, the new model does not fit the predictions well. Nonetheless, with over 60-percent of the variation explained by the model, it is possible to draw some conclusions. Figure 7 provides an additional visualization resource for some patterns explored here.

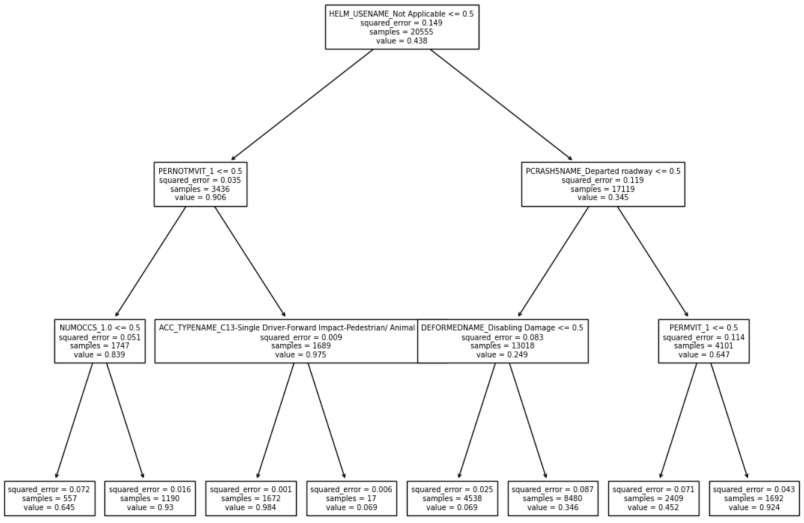

Figure 7: Global Surrogate Model

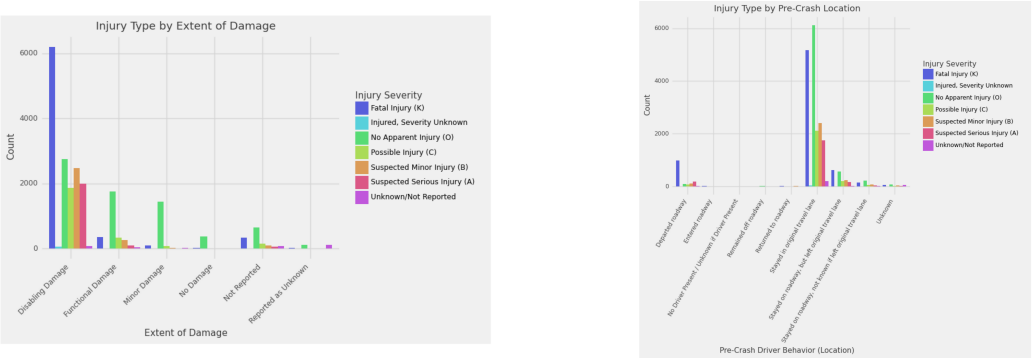

Figure 8: Damage Type and Pre-Crash Location

The surrogate model's narrative is clear and provides a few interesting and interpretable

stories from the data. In the most predictive case, the model predicts with 98.4-percent accuracy that a single pedestrian struck by a vehicle will be killed.

5 Discussion

To start, limited data restricted the utility of visualizations. This limitation takes the form of a 'near-base rate fallacy,' and is best explained through an example. Suppose an individual is asked whether it is more likely to see a fatality occur on a sunny day or a rainy day. It's reasonable for that individual to assume rainy due to substantial evidence that rain represents a more 'dangerous' condition. However, since there are likely three times as many sunny days, the vast majority of traffic fatalities occur on sunny days. While this problem is easily accounted for by statistical learning techniques, without proper documentation of hyper-localized weather patterns, this data provides no inherent ability to counter the issue.

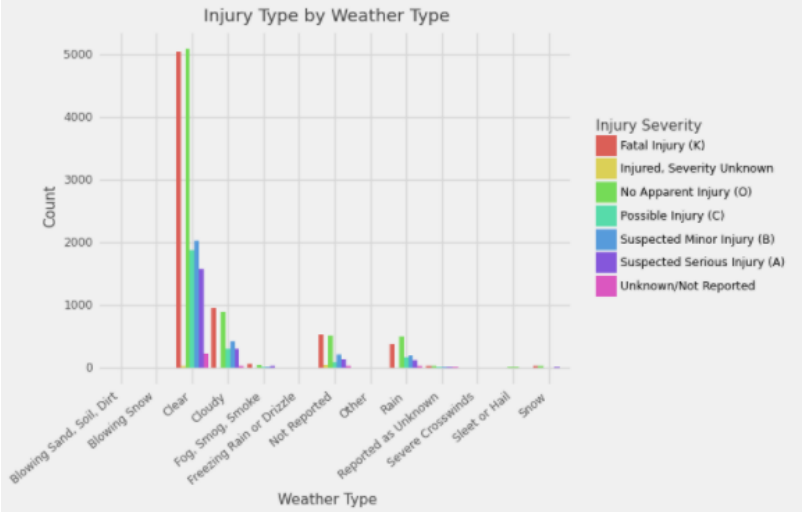

Figure 9: Fatalities by Weather Condition

The same fault occurs in other variables as well, with Figure X pulling a few notable cases. In all plots, it is likely that each spike correlates to the system with the most vehicle miles traveled.

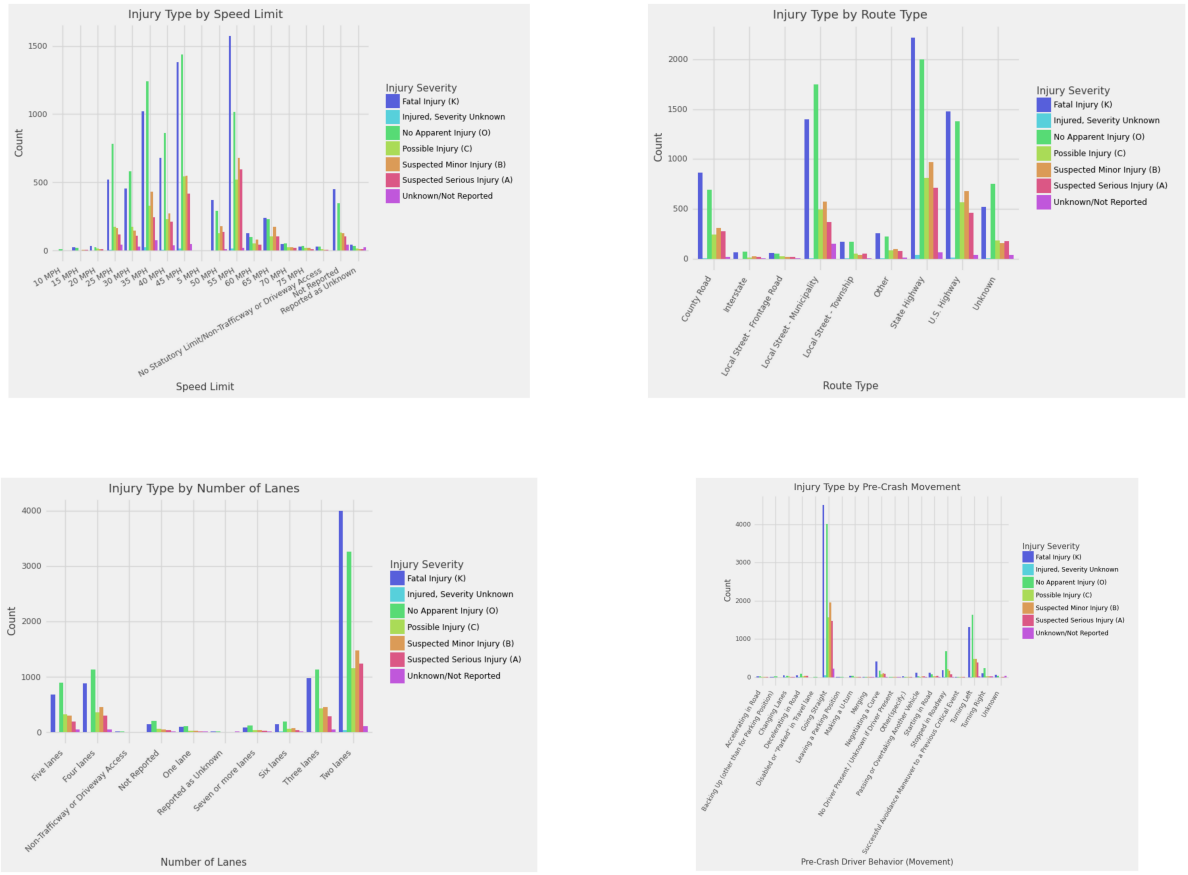

Figure 10: Near-Base Rate Fallacy Demonstration

With respect to future, policy-relevant research, one particular finding stood out. Figure 11 summarizes the result. Here we see that as the number of passengers increases, the model becomes less likely to predict a fatality. Further, in a similar model constructed where each number of passengers takes on a dummy value, the results reiterate the same effect but highlight a substantial increase in predictive accuracy of a fatality when the one-occupant variable is activated. This is then followed by a series of comparative decreases in the likelihood of predicting a fatality when two, three, and four occupants are then each respectively activated.

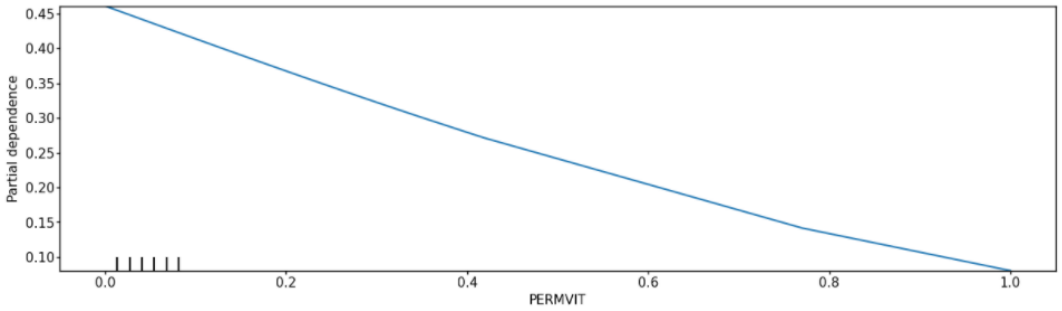

Figure 11: Fatality Prediction by the Number of Motor Vehicle Occupants (Continuous)

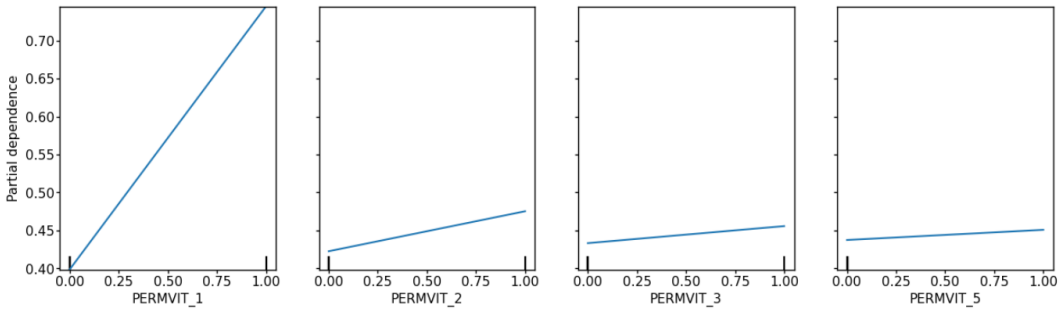

Figure 12: Fatality Prediction by the Number of Motor Vehicle Occupants (Dummy)

This result deserves particular attention due to the existence of graduated driver licensing systems designed to require a multiple phase approach to fully legalized, adolescent driving. Such programs exist in all 50 states and DC, with many states limiting the number of passengers a new driver can have in their vehicle. These programs have been well studied, and generally find sizable effects in the form of reduced crash rates of young drivers.[12] To investigate this further, this project tested the potential interaction between one-passenger-vehicles and the age of the driver on traffic fatalities. Figures 13 and 14 display these results and appear to demonstrate that the effect does not significantly differ by age. Instead, the plots illustrate that across all age cohorts, holding only one occupant increases the model's likelihood of predicting a fatality.

[12]Motao Zhu, Songzhu Zhao, Leann Long, Allison Curry, "Association of Graduated Driver Liscensing With Driver, Non-Driver, and Total Fatalities Among Adolescents," *American Journal of Preventive Medicine* 51, no. 1 (July 2016): 1, `https://doi.org/10.1016/j.amepre.2016.02.024`.

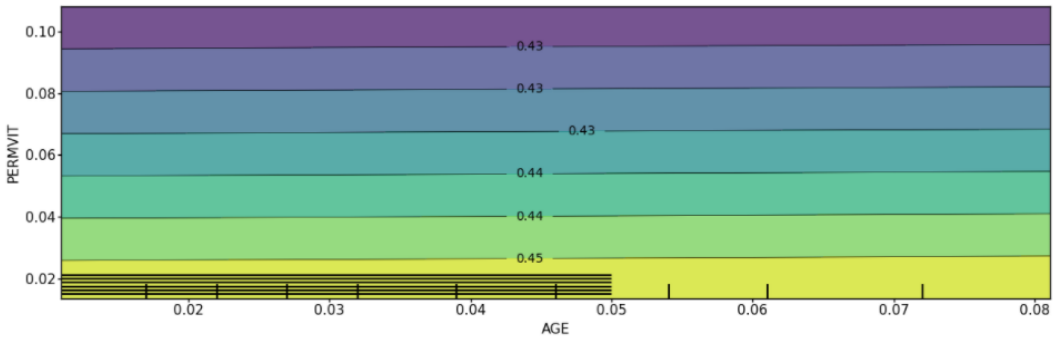

Figure 13: Fatality Prediction by the Number of Motor Vehicle Occupants and Age

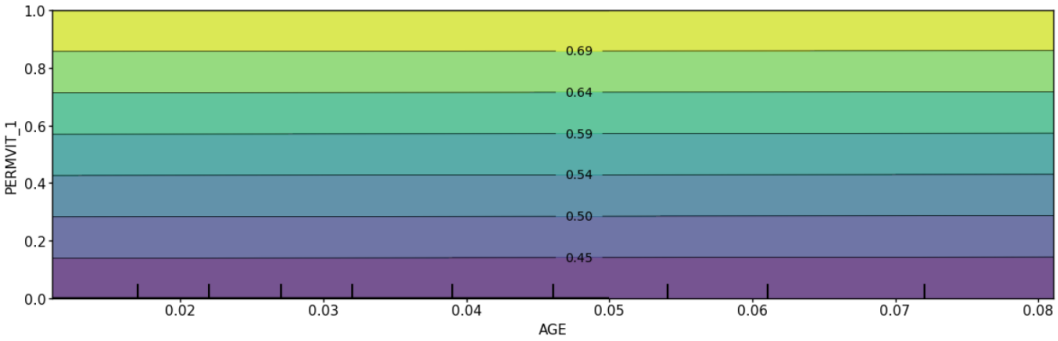

Figure 14: Fatality Prediction by the One Motor Vehicle Occupants and Age

Despite consistent results, further evidence must be established in order to provide empirically-based policy recommendations. One possible explanation stems from the model's data, which does not account for total accidents and instead predicts based on which passenger is likely to be killed from a pool of total traffic fatalities. Conceptually, it is plausible that a second (or third, etc.) set of eyes provides increase protection against incoming dangers. This rationale similarly fuels the potential narrative illustrated by Figure 7. Finally, additional insight may be found by exploring other years with the same predictive modeling approach.[13]

[13]This project attempted to run the 2008 data through the same model. Please see the public repository for (limited) results.

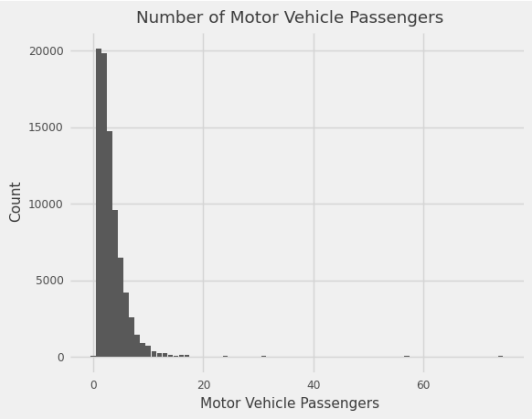

Figure 15: Total observations by Number of Motor Vehicle Occupants