US Traffic Fatalities
A Data Science Analysis

Kirk Hazen

Introduction

Despite improvements in vehicle safety since 1950, there were nearly 36,082 deaths on US roads in 2019. The Fatality Analysis Reporting System (FARS) of the National Highway Traffic Safety Administration (NHTSA) was established to provide data on these deaths in the hopes of finding causes in the patterns. This report provides an analysis towards those ends.

The motivation

As the FARS manual notes, "Crashes each year result in thousands of lives lost, hundreds of thousands of injured victims, and billions of dollars in property damage. Accurate data are required to support the development, implementation, and assessment of highway safety programs aimed at reducing this toll." The FARS data were collected to help improve traveler safety, and proper analysis and machine learning models can help us move towards safer travel.

Specifying the data

The FARS data are specific in that "To qualify as a FARS case, the crash had to involve a motor vehicle traveling on a trafficway customarily open to the public, and must have resulted in the death of a motorist or a non-motorist within 30 days of the crash." This dataset does not include crashes where a person did not die. Accordingly, this dataset cannot be used to assess which factors led to non-fatal results from a crash. The data were chosen from the last pre-pandemic year of driving, 2019, and can be accessed from NHTSA: https://www.nhtsa.gov/.

Data

About the data

These data are entered in at crash sites and assigned state case numbers by the managers of the FARS dataset. For numerous fields, including weather conditions, objects impacted by crashes, influence of drugs or alcohol, the data may include ‘unknown’. In some ways this limits the potential of the analysis but overall these cases are not numerous enough to negatively impact the final analysis.

There are several data tables available for analysis, including those that focus on state cases (one per accident), vehicles, people (including pedestrians), factors possibly contributing to the crash, vision, race, and drugs.

After wrangling many of these tables and exploring how possible joins might enhance the explanatory power of predictive models, it was determined that the accident table with 91 columns and 33,487 rows contained the crucial information for the accident.

Wrangling the data

There were several issues with the dataset that required changes.

1. Many of the columns had integers to represent categorical information

    As a solution for columns where there were also adjacent columns with names pertaining to the same categories (Column Day has 1 and Column Day_Name has Sunday), the column with the integer categorization was dropped. The exception was DRUNK_DR where it was a binary number (1 or 0) which was changed to a categorical variable.

2. Not all the columns in the dataset appear to be relevant to assessing predictions of fatalities.

    Columns such as the name of the hour of arrival to the hospital were winnowed down.

3. Larger groupings were created in new columns to explore possible patterns.

    These include region (based on Census Bureau designations), Weekend/Weekday, Day/Night. Averages for each of these were also calculated.

4. Fatalities did not fluctuate much throughout the dataset.

    The basic premise of the dataset, that every row has at least 1 fatality, became a statistical hindrance for analysis. The mean for fatalities was only 1.1, and the standard deviation was only 0.5. In many ways, having

a lower number of fatalities per accident is good news, but it does not permit much insight analysis.

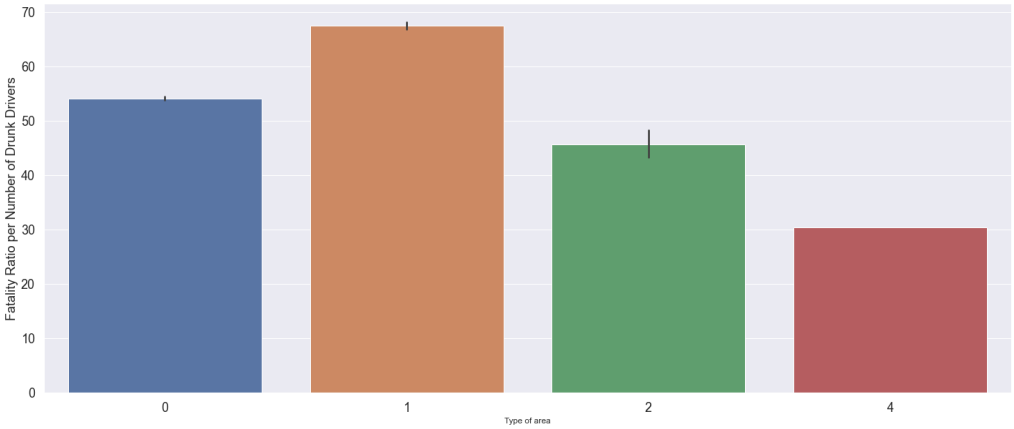
To correct for this issue, a ratio was created in a calculated column to explore the number of fatalities in relation to the number of people (including pedestrians) involved in the accident. The Fatality Ratio allows for a better assessment for the loss of human life in these accidents.

Column	Mean	Standard Deviation
Fatalities	1.1	0.5
Fatality Ratio	57.4	28.5

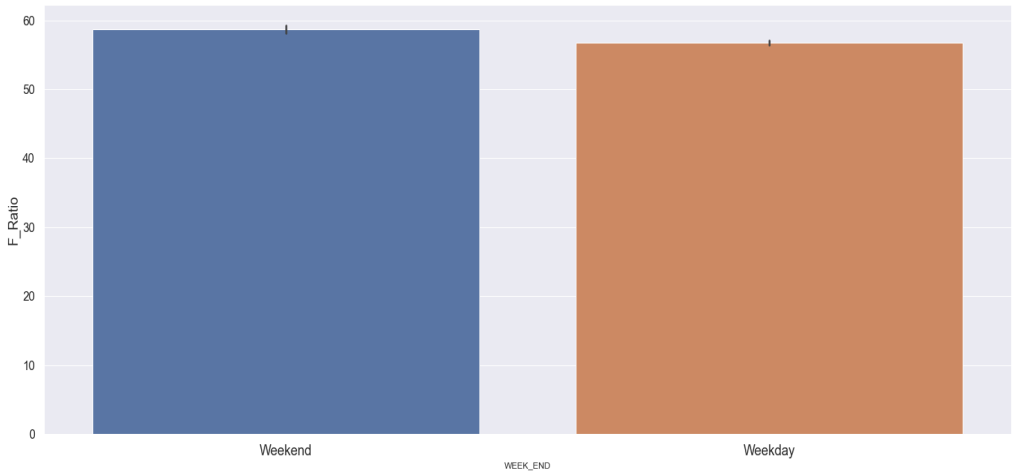
Fatality per Accident	Count	Percent of Fatality Share
1	30895	85.62%
2	1967	5.45%
3	286	0.79%
4	69	0.19%
5	12	0.03%
6	5	0.01%
7	3	0.01%
8	1	0.003%

Exploratory Data Analysis

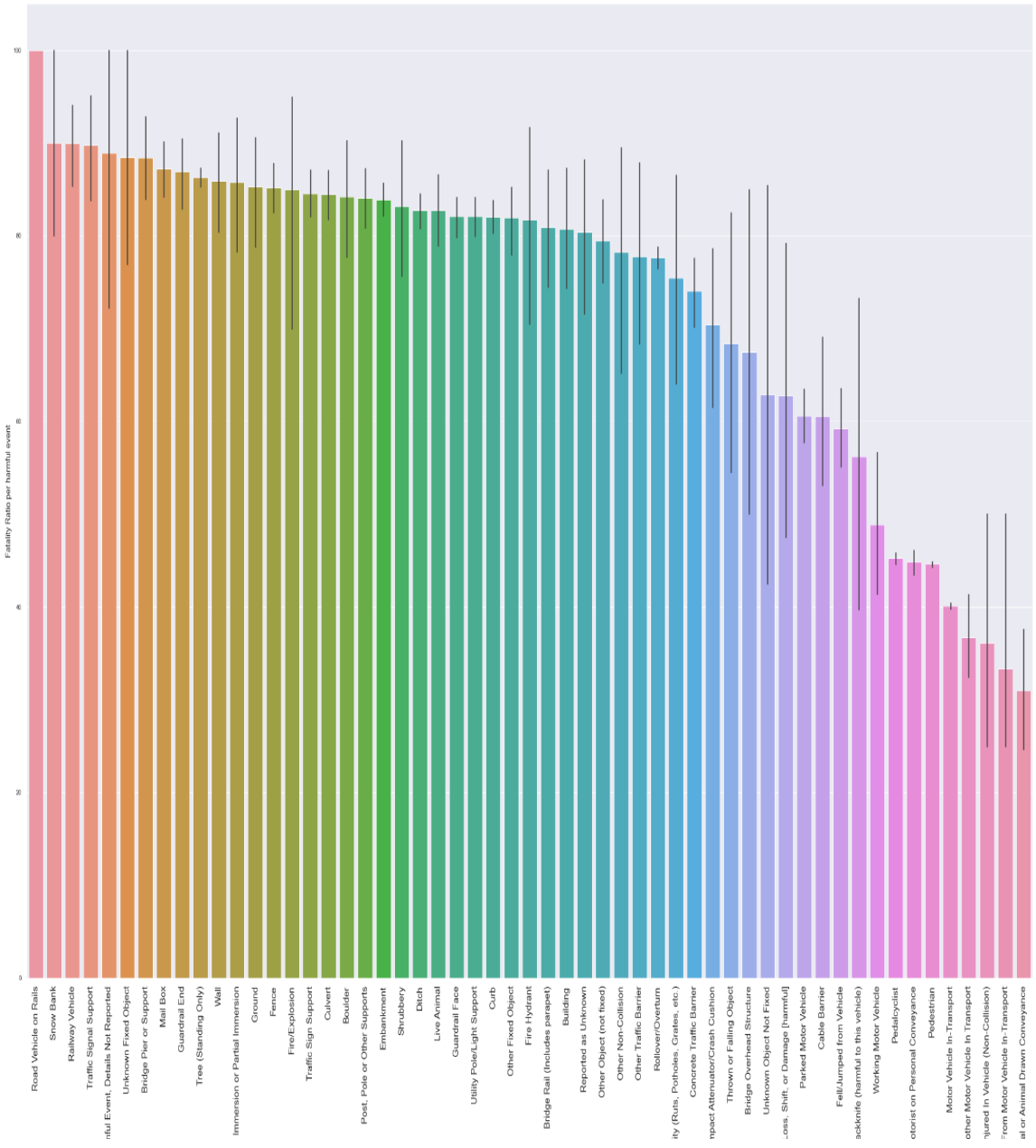
In the EDA, there is variability in Fatality Ratios for potentially influential features.



Other assumptions about potential variability, such as between weekday and weekend, did not demonstrate much variation.



This variability was especially true across the 55 different types of harmful events.



Findings

The machine learning models were assessed based on Mean Absolute Error and Root Mean Squared Error. Using cross validation, the set of possible features, 179 after transforming categorical variables, were paired down to 13 features which had a top four ranking in the Recursive Feature Elimination assessment.

Category	Ranking from RFE with Lasso CV
PEDS	Rank 1.000
PERSONS	Rank 1.000
DRUNK_DR	Rank 1.000
RUR_URBNAME_Rural	Rank 1.000
HARM_EVNAME_Motor Vehicle In-Transport	Rank 1.000
HARM_EVNAME_Non-Motorist on Personal Conveyance	Rank 1.000
HARM_EVNAME_Pedalcyclist	Rank 1.000
HARM_EVNAME_Pedestrian	Rank 1.000
HARM_EVNAME_Traffic Signal Support	Rank 1.000
HARM_EVNAME_Tree (Standing Only)	Rank 1.000
HARM_EVNAME_Other Object (not fixed)	Rank 2.000
HARM_EVNAME_Other Non-Collision	Rank 3.000
HARM_EVNAME_Other Fixed Object	Rank 4.000

One of the original research questions was whether or not there existed a difference in fatalities for rural and urban areas. Rural was chosen as a top ranked factor for contributing to a higher fatality ratio. Despite the higher population density in urban areas and thus higher traffic density with more objects to hit, rural areas are still a top ranked feature. Drunk driving and several harmful events are also top ranked.