**Contributing Factors for US Traffic Fatalities**

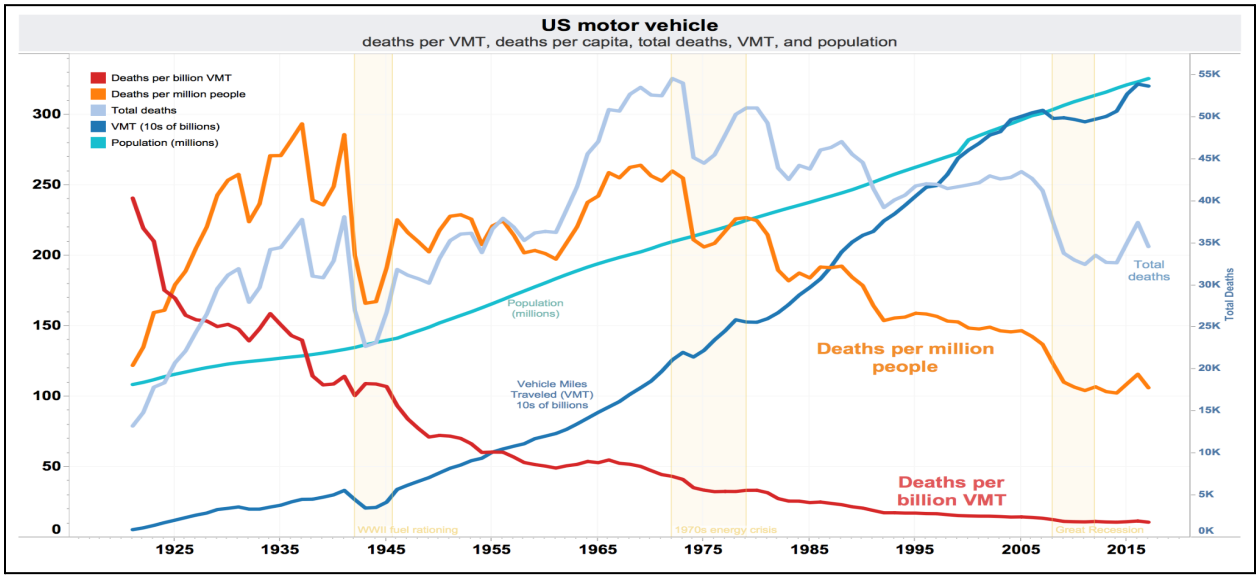
Kirk Hazen
Springboard
Data Science Career Track
Raghunandan Patthar, Mentor
Sep 21, 2022

ORCID: https://orcid.org/0000-0001-6402-5176
Acknowledgements: I would like to thank my mentor, Raghunandan Patthar, for guidance with this project, and my on-call mentor Wei Ang for statistical and machine learning advice.
Email: Kirk.Hazen@icloud.com

# Contributing Factors for US Traffic Fatalities

Across both rural and urban areas in the US, traffic fatalities persist despite vast improvement in automobile safety. As can be seen in Figure 1, the ratio of deaths has decreased over the last century, but every year thousands of people needlessly die in traffic accidents. If machine learning models can predict fatalities given certain criteria, we can better assess what factors contribute to traffic fatalities. In addition, we can understand the impact that will result from the continuing transformation of rural areas into more populated communities.

Figure 1:




By Dennis Bratland - Own work, CC BY-SA 4.0

Data was drawn from the Fatality Analysis Reporting System (FARS) of the National Highway Traffic Safety Administration (NHTSA). 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."

## Methods

### Data

The data will come from US states and look primarily at the differences in many factors coinciding with fatal accidents. Every row of the data contains at least 1 fatality, so this dataset is not a collection of all accidents. For some of the feature types, the scope is limited by the divides in the categories. For example, the dataset only has a binary divide in rurality along with unknowns, although the reality is more complex. For this Capstone 2 project, a time-series analysis is not the goal. Considering the impact of the pandemic, the year 2019 was chosen as the last “normal” year where data is available. The FARS data was collected to help improve traveler

safety. 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."[1]

**Data Wrangling Approach**

After casting the data to utf-8 format, the accident table was imported as a pandas DataFrame and has 33,487 rows across 91 columns. Initial data wrangling reduced the number of features (such as state case number) which could not influence fatalities in an accident. For this analysis, several types of linear regression were applied to the data to find the best functioning model. In addition, at least initially, rural/urban, drunk driving, weather, light condition, work zone, and pedestrian columns all appear to be important for this analysis. Some are as 'object' currently and some are as integers; properly adjusting each as either an integer or as a multi-level factor will be important for the analysis.

**Exploratory Data Analysis**

In the exploratory data analysis phase, many of the potential features were found to show minimal variability. The initial target measure was the number of fatalities but as can be seen in Table 1, the overwhelming majority of accidents only have one fatality. The total was 36,082 fatalities, and signal fatality accidents were 85.62% of all accidents in 2019.

Table 1: Number of Fatalities per Accident

| 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% |

With this quirk to the dataset, an alternative measure needed to be found. Averages based on divisions of region, day/night, weekday/weekend, and rural/urban were assessed, but in the end a measure was devised for the fatality ratio. The fatality ratio is the total number of fatalities for an

[1] National Center for Statistics and Analysis. (2022, March). *Fatality Analysis Reporting System analytical user's manual*, 1975-2020 (Report No. DOT HS 813 254). National Highway Traffic Safety Administration.

accident divided by the number of people in the accident (including pedestrians) and then multiplied by a 100 in order to make a range between 1-100.

**Outliers vs Typical Cases**

Some cases were unusual. There was a crash in Pennsylvania where 59 vehicles were involved, which resulted in 2 fatalities. In part the causes were blowing snow and a jackknifed truck. This crash is an outlier in how low its fatality ratio is. Some crashes stretched the numbers in other directions. A crash with eight fatalities happened in Mississippi. It involved two vehicles where one vehicle had nine people and the other had one person. Three instances of fatality ratios below five, where 1 person died with more than 20 people involved, were eventually excluded because of how they displaced axes on graphs.

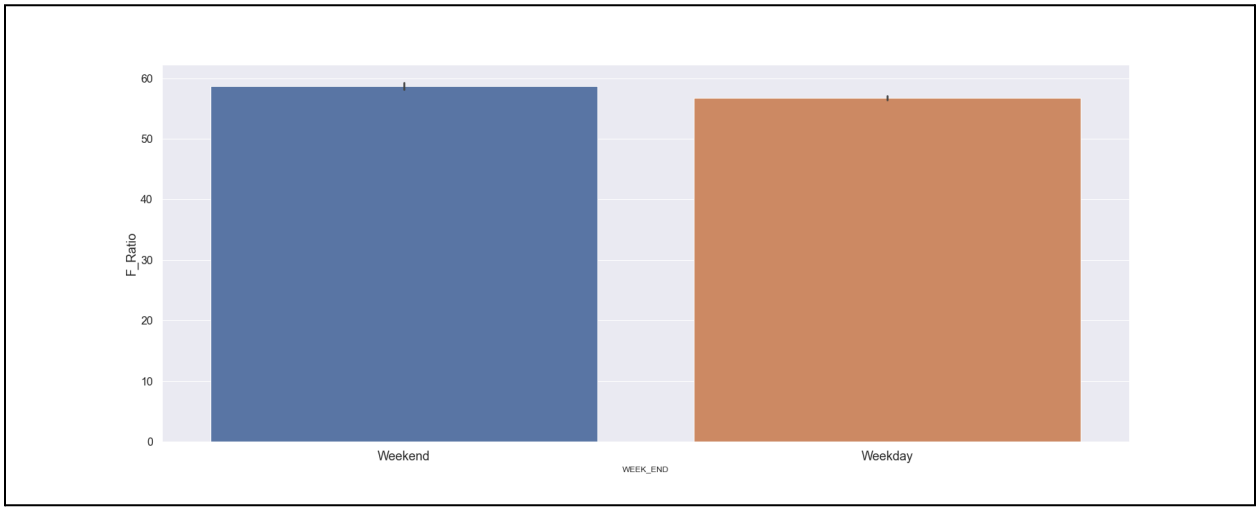
## Results

**Differences in Factors**

After exploratory data analysis, the number of drunk drivers, weather conditions, harmful events, and day/night appeared to be the leading features. Rural and urban differences did not appear to be sizable enough to justify testing them as separate datasets, but the factor of Rural vs. Urban was kept for the total dataset.

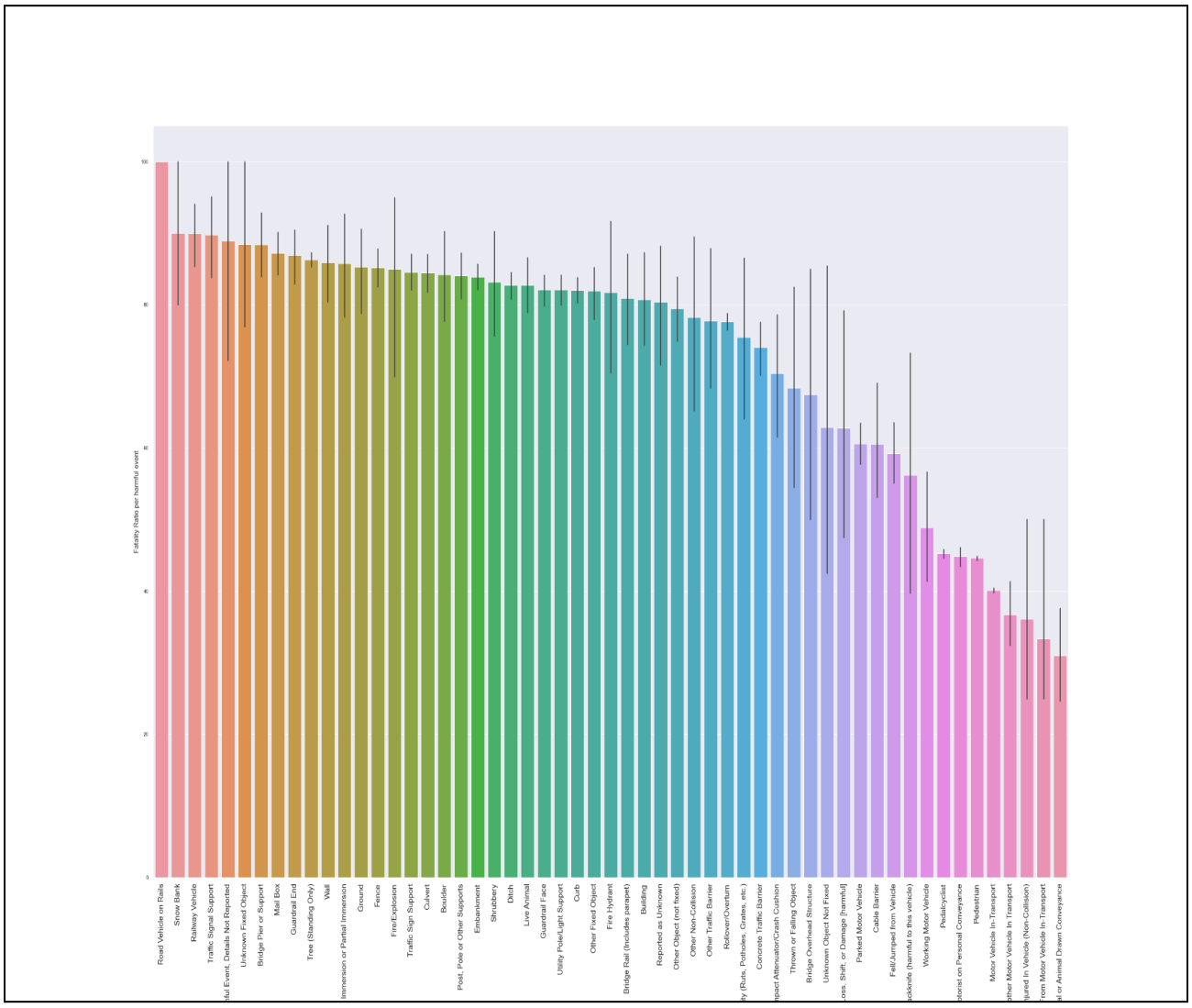
Like most features, in making initial hypotheses about what might contribute to a higher fatality ratio, the day of the week was considered a possible influencing factor, with the assumption being that weekends might see a higher fatality ratio. This assumption turns out to be true for the 2019 FARS data, but this contrast between weekend and weekday was not selected as a top-ranked feature in the machine learning models.

Figure 2: Fatality Ratio for Weekend vs. Weekday



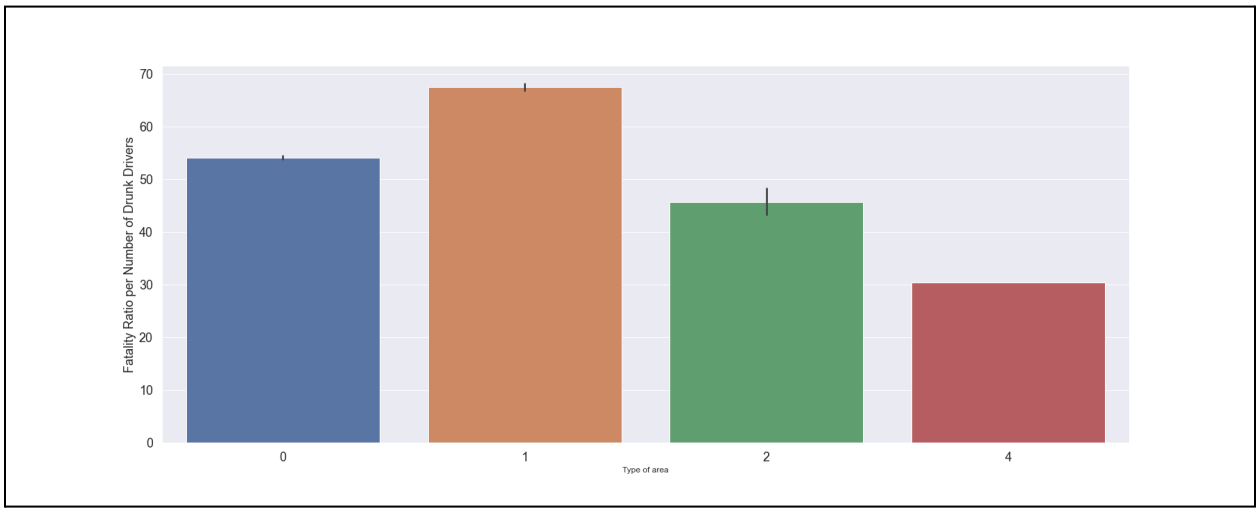
There are many contributing features and only some of them are considered top ranked for influencing the fatality ratio. Even within the column of harmful events, with 55 different categories, some of the factors that have the highest fatality ratio, such as “Road Vehicle on Rails” (i.e. getting hit by a train) and “snow bank” were not selected as influential most likely because they are not common occurrences for accidents overall. Wide variability in fatality ratios does exist for this column overall and these differences are noted in the number of high ranked harmful events in the best machine learning models. Figure 3 is presented simply to illustrate the variability. 55 bins on the x-axis does not allow for readability.

Figure 3: Fatality ratio by the type of harmful event



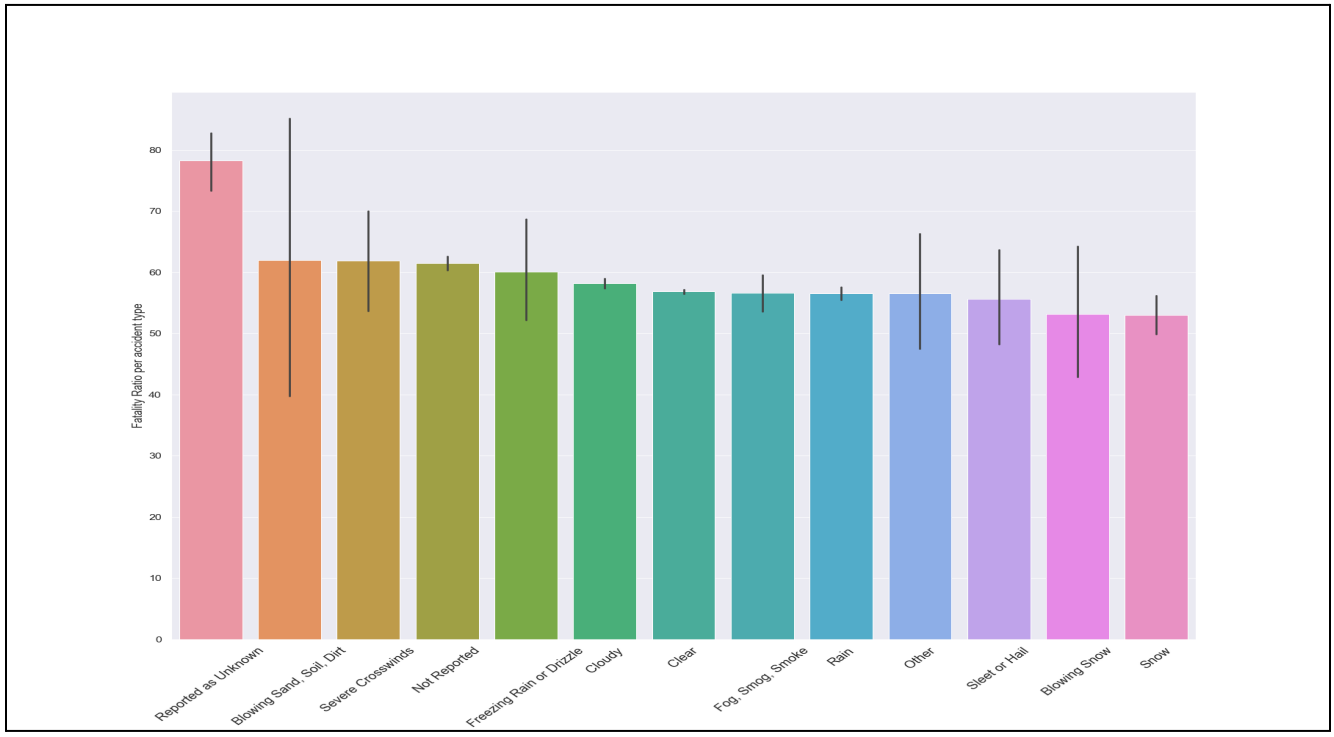
The number of drunk drivers was also chosen as a highly ranked factor for the fatality ratio. In 2019, only one accident had more than two drunk drivers. In Figure 4, accidents with just one drunk driver were most common and also had the highest fatality ratio.

Figure 4: Fatality ratio by number of drunk drivers



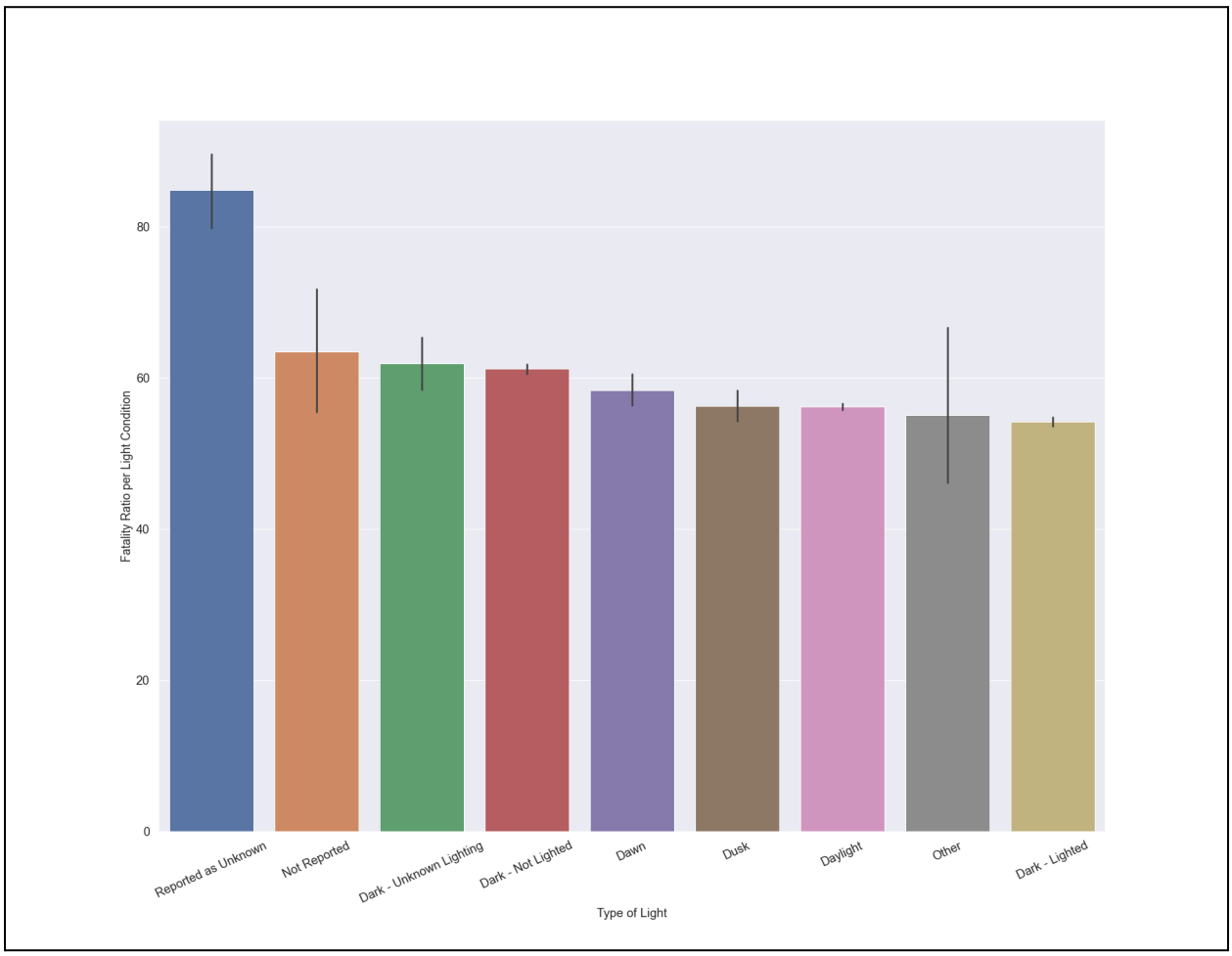

Despite an initial hypothesis about weather contributing to fatalities, the fatality ratio did not fluctuate much between clear and other conditions. The only odd-ball category is ‘unknown’ which did have a much higher fatality ratio.

Figure 5: Fatality ratio by weather condition



Additionally, it was assumed that certain light conditions would also contribute to a higher fatality ratio ratio, but like the findings from the weather category, there was not much variation in known light conditions nor was there any clear pattern pertaining to darkness and light.

Figure 6: Fatality ratio by light condition




**Reducing Possible Factors**

Through Recursive Feature Elimination with Lasso, I was able to trim the features to those of greatest importance and get a model that has decent metrics without overfitting. As can be seen in Table 2, several models focusing on the linear fatality ratio as the target were attempted, including Ordinary Least Squares, Linear Regression, Lasso, and Decision Tree Regression. The last approach uses 'standard deviation reduction' in order to assess when there is a "decrease in standard deviation after a dataset is split on an attribute. Constructing a decision tree is all about

finding attributes that return the highest standard deviation reduction (i.e., the most homogeneous branches)"(Sayad 2022).

Table 2: Models and their metrics

| Feature Set | Model | Mean Absolute Error | Root Mean Squared Error | Coefficient of determination- $R^2$ |
|---|---|---|---|---|
| Full | Ordinary Least Squares | 57.47 | 59.5 | -3.35 |
| Full | Lasso (alpha = 1.0) | 12.22 | 15.63 | 0.7 |
| Full | Linear Regression | 10.99 | 15.21 | 0.72 |
| Full | Decision Tree Regressor max depth = 5 | 4.24 | 9.91 | 0.88 |
| Trimmed | Lasso (alpha = 1.0) | 12.22 | 15.63 | 0.7 |
| Trimmed | Linear Regression | 11.15 | 15.3 | 0.71 |
| Trimmed | Decision Tree Regressor max depth = 5 | 4.25 | 9.92 | 0.88 |

Max depth was determined through models with lower and higher max depths, and those in turn returned worse metrics (both higher MAE and RMSE scores). The Decision Tree Regressor was the best model in terms of MAE, RMSE, and $R^2$ with the trimmed data set. From assessment of Lasso regression using cross-fold validation, 179 features were given ranks of importance, and those features with ranks below five were chosen as the most important features.

- Number of pedestrians
- Number of persons in vehicles
- Number of drunk drivers
- Rural areas
- Harmful events:
    - crash with a motor vehicle in-transport
    - crash with motorist on personal conveyance
    - crash with a cyclist
    - crash with a pedestrian
    - crash with a traffic signal support
    - Crash with standing tree
    - Crash with other not-fixed object
    - Crash with fixed object

- Wreck without crash (e.g. rollover)

In all 13 factors were found to best account for higher fatality ratios. Many of these are harmful events where the motorist crashed into something. Although the category of other fixed objects is slightly lower ranked, the category of non-fixed objects is more highly ranked, which begs many questions as to what those objects were. Crashing into a tree, a traffic light pole, a pedestrian, a cyclist, or another vehicle were also top ranked for a higher fatality ratio. Because all people involved in an accident were included in the fatality ratio, it does account for pedestrians also. The number of pedestrians and persons in a vehicle also contributes to the fatality ratio. It should be noted however that some cases with larger numbers of people involved, like the one Pennsylvania accident involving 59 people with two fatalities, actually have much lower fatality ratios. Drunk driving and rurality both are ranked as top contributors to the model predicting the fatality ratio. Drunk driving is a simple count of the number of drunk drivers in the vehicles involved in the crash, and its contributions can be easily deduced.

Why rurality is also a top-ranked factor invites many questions: Roads may have less lighting in rural areas, but if so, why did light conditions not appear as a factor? The population density is lower in rural areas, so does this factor contribute to higher speeds? Do the road conditions themselves in rural areas contribute to more deadly crashes? Rurality itself is not a cause of a higher fatality ratio, but instead it is a set of conditions that foster the types of accidents where more people die. Understanding those conditions is an important ongoing investigation.

## Discussion

At the start of this process, the focus was on the trend for urban areas to grow and rural areas to shrink over the last century and its continuation for the foreseeable future. With the growth in urban areas, it was assumed that there might be an increase in the relative number of traffic fatalities. By using machine learning models and the many contributing factors of the FARS data, 13 factors were found to be the most influential. Drunk driving has been a widely known issue for decades, and organizations such as Mothers Against Drunk Driving (https://madd.org/) have held public awareness campaigns and lobbied for years on behalf of families who have lost loved ones to this ultimately preventable cause of death. Other top-ranked factors could be mitigated through greater use of technology such as Automatic Emergency Braking systems. These systems can stop the vehicle quickly when cameras/radar detect objects (such as light poles or other vehicles) in the driver's path. These systems have been enhanced so that Automatic Emergency Braking With Pedestrian Detection is becoming the standard for vehicles (Barry 2022).

Although the full dataset contains many other possible factors, by leveraging the power of machine learning, the current model allows for relatively accurate prediction with a low error rate and provides the opportunity to focus on the factors that contribute most to the fatality ratio.