Using Machine Learning to Identify Factors Contributing to Higher Fatalities in Florida Traffic Crashes

Sarah Thapa & Rafiq Islam

FSU | FLORIDA STATE UNIVERSITY

FSU UNDERGRADUATE RESEARCH OPPORTUNITY PROGRAM
CENTER FOR UNDERGRADUATE RESEARCH & ACADEMIC ENGAGEMENT

Introduction

Traffic safety has long been a critical component of public health and transportation policy across the United States. Among the states, Florida has developed a particularly concerning reputation for high crash rates and risky driving behavior. According to the Florida Department of Highway Safety and Motor Vehicles (2024), the state recorded 4,814 alcohol-related crashes, 271 of which were fatal. These statistics underscore the urgent need to examine the underlying causes of traffic fatalities and to implement evidence-based interventions aimed at reducing them. Addressing this issue requires a systematic understanding of the various factors that contribute to fatal crashes, including driver behavior, environmental conditions, and roadway design.

With the advancements in data collection and analytic technologies, researchers now have access to traffic datasets that enable predictive analyses of crash patterns. Applying these tools can help identify the most significant contributors to fatal incidents, allowing policymakers and law enforcement agencies to design targeted safety initiatives. Furthermore, by translating these findings into public education and awareness campaigns, the general population can become better equipped to adopt safer driving practices.

Our Question

What are the significant factors contributing to the higher fatalities in Florida Traffic?

Importance of this Research

Public safety is critical for any community. Having access to accurate data and thorough analysis enables researchers and policymakers to identify the underlying causes behind the growing number of roadway fatalities. The numbers do not lie!

Having reliable statistical evidence highlighting the most significant factors contributing to fatal crashes provides valuable insight for both the public and decision-makers. This knowledge can be used to develop informed policies, implement effective safety measures, and ultimately create safer roads, cities, and communities for everyone.

Machine Learning Models

The two Machine Learning (ML) Models we ended up going with include:

1. **Logistic Regression**: Used for creating a yes or no outcome based on given factors
2. **Random Forest**: Combines smaller nodes called decision trees to make one final prediction

Using these two machine learning models, we can develop a predictive framework to address our research question. The models will utilize data from the Fatality Analysis Reporting System (FARS), which has been thoroughly cleaned and refined through multiple preprocessing steps and exploratory data analysis (EDA) to generate accurate predictions.

Method

The first step we took was to identify sources from which we could collect data. This initial stage was crucial because the success of our analysis depended on the quality and scope of the information we gathered. It is important to note that, in order to properly implement the two models we plan to use, we needed a sufficiently large dataset. Additionally, we require a dataset capable of addressing our research question: ***What are the significant factors contributing to the higher fatalities in Florida traffic crashes?***

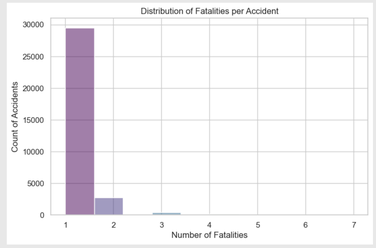
We ultimately chose to use traffic data from the **National Highway Traffic Safety Administration** (NHTSA). The NHTSA is a federal agency under the U.S. Department of Transportation that prioritizes roadway safety by enforcing national standards (Federal Register, n.d.). To support these efforts, the NHTSA collects data from the **Fatality Analysis Reporting System** (FARS). This dataset provided the level of depth and reliability we needed to explore the causes of severe traffic accidents in Florida.

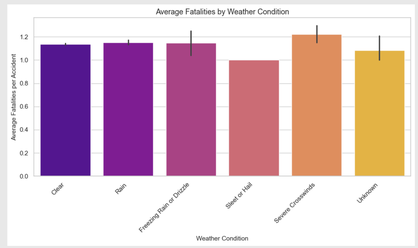
After retrieving and cleaning the data, we ended up with a dataset containing **33,014 variables** to observe and analyze. During this stage, we removed missing or irrelevant values and reformatted inconsistent data entries to maintain uniformity. The columns we investigated are **road area type, road surface, manner of collision, time of day, day of week, weather condition, light condition, speeding involvement, driver blood alcohol level, road surface, road surface drowsy driver, vehicle type, rollover involvement, license status, and drivers age**. Before starting logistic regression, we had to standardize the data, which ultimately means we converted the raw data into a unified format for consciences.

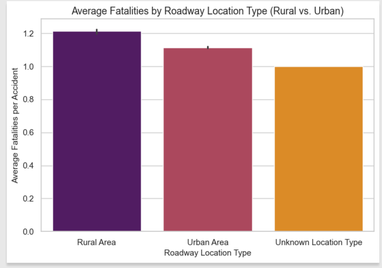
Exploratory Data Analysis (EDA)

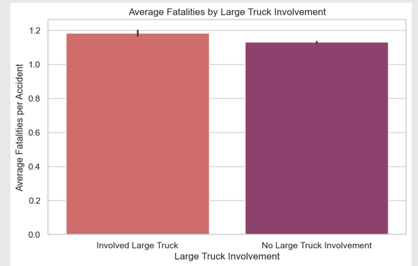
After cleaning the data, the next critical step before applying machine learning models is to conduct an exploratory data analysis (EDA). EDA provides a comprehensive understanding of the dataset, helping to identify patterns and relationships that may be meaningful or spurious. With the cleaned data, we can focus on variables that are truly significant rather than those that simply exhibit correlation. During this process, we generated visualizations including histograms, pie charts, and bar charts to summarize and interpret the data. The table below presents the first chart, which demonstrates that every crash in the dataset resulted in at least one fatality. This visualization highlights that our analysis is focused on identifying the factors associated with the highest number of deaths per crash.

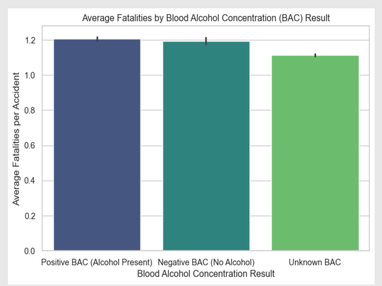
Data Based on the Accident sheet (2018-2022)

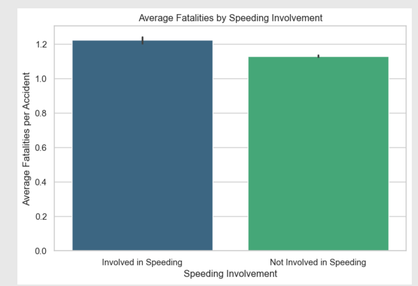




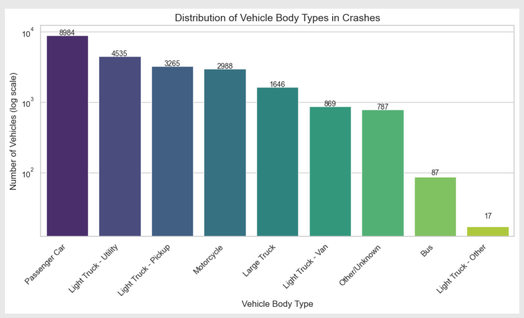






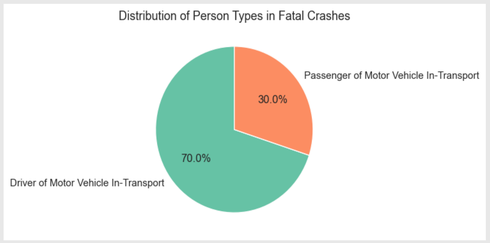


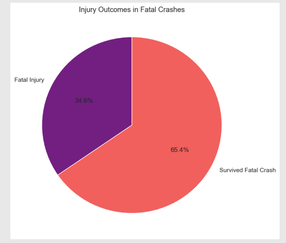
Data based on the Vehicle sheet (2018-2022)



Exploratory Data Analysis (EDA continued)

Data based on the Person sheet (2018-2022)





Limitations

- Some data may be missing. For example, some of the location type for crashes are unknown
- The dataset is reported by law enforcement which means there could be instances that are not on record
- Some collection of data over the years changed so there are values that are missing and automatically taken out of potential factors

Machine Learning Results

Logistic Regression: Predictive performance

The logistic regression model achieved **AUC = 0.939**

Confusion matrix (test set):
- True Negative = 5,597
- False Positive = 857
- False Negative = 393
- True Positive = 3,021

Key classification metrics:
- Precision: 0.78
- Recall (Sensitivity): 0.88
- F1-score: 0.83

Overall accuracy: 0.87, with balanced performance across classes

Key factors associated with decreased fatal injury risk

Several categories were strongly protective relative to baselines:
- **Alcohol Test = 3.0** (Not tested): OR ≈ **0.007**
- **Road surface = 2** (Dry Road): OR ≈ **0.17** (protective vs baseline)
- **No rollover = 2** (No Rollover): OR ≈ **0.41**

Random Forest permutation importance

The top contributors to predictive Permutation performance were:
Alcohol Test= 3.0 (Not tested)
Road surface= 2 (Dry Road)
Person Type= 2.0 (Occupant)
Alcohol Test= 5.0 (Unknown if tested)
Vehicle body type = 7 (Motorcycle)
Restraint Use = 2.0 (No seatbelt used)

Summary of findings

Based on both ML results, fatal injury risk was most strongly associated with:
- **Alcohol-related indicators**
- **Restraint use**
- **Vehicle body type**
- **Collision manner**
- **Road surface condition**
- **Occupant position**
- **Age**

The strong cross-model agreement suggests these factors are robust predictors in this dataset.

References

Federal Register. (n.d.). *National Highway Traffic Safety Administration*. U.S. Government Publishing Office. Retrieved February 19, 2026, from https://www.federalregister.gov/agencies/national-highway-traffic-safety-administration.

Florida Department of Highway Safety and Motor Vehicles. (2024, February). *By the numbers: Traffic statistics summary* — February 2024. https://www.flhsmv.gov/pdf/opengov/by-the-numbers_feb-24.pdf

Kavlakoglu, E. (n.d.). *What is random forest?* IBM. https://www.ibm.com/think/topics/random-forest

Lee, F. (n.d.). *What is logistic regression?* IBM. https://www.ibm.com/think/topics/logistic-regression.

National Center for Statistics and Analysis. (2025, August). *Fatality Analysis Reporting System analytical user's manual*, 1975-2023 (Report No. DOT HS 813 706). National Highway Traffic Safety Administration. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813706.

National Center for Statistics and Analysis. (2025). *Fatality Analysis Reporting System analytical user's manual, 1975–2023* (Report No. DOT HS 813 706). National Highway Traffic Safety Administration. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813706.

National Highway Traffic Safety Administration. (2019, October). *Fatality Analysis Reporting System (FARS) auxiliary datasets analytical user's manual, 1982–2018* (Report No. DOT HS 812 829). U.S. Department of Transportation. https://www.safehomealabama.gov/wp-content/uploads/2019/11/Fatality-Analysis-Reporting-System-FARS-Auxiliary-Datasets-Analytical-Users-Manual-1982-2018.pdf.

Roselli, M., & McNelis, J. (2014). *Florida car crashes 2014*. Roselli & McNelis, P.A. https://www.rosellimcnelis.com/florida-car-crashes-2014/.

https://github.com/saraht16/floridaTrafficFatReport.git