Analysis of Factors Contributing to Fatality in U.S. Heavy Duty Truck Crashes

Philip M. Situmorang
Volpe Center
philip.situmorang@dot.gov

Abstract

In the United States, there were roughly 4,000 fatal crashes involving heavy duty trucks per year between 2019 and 2021, a rise from roughly 3,000 fatal crashes in 2010. To reduce the loss of lives, the United States Congress commissioned the Federal Motor Carrier Safety Administration (FMCSA) to study factors which contribute to fatal crashes involving motor carriers through the Consolidated Approriations Act of 2021. In that same year, the FMCSA launched the Crash Causal Factors Program – Heavy Duty Truck Study (CCFP-HDTS), which seeks to identify factors contributing to fatal crashes among heavy duty trucks and inform countermeasures. Currently, the study seeks to analyze over 70 factors from future sampling of crashes involving heavy duty trucks, defined as trucks with gross vehicle weight rating exceeding 26,000 pounds. This research uses data immediately available from FMCSA's Motor Carrier Management Information System (MCMIS) to answer a subset of the research questions outlined. The purpose of this analysis is to serve as a pilot study and framework for the future analysis of CCFP-HDTS sample data.

1 Introduction

Risk levels of a vehicle accident are generally estimated by the probability of its occurrence and the severity of its outcome. Likewise, efforts to lower accident risk on the road are typically aimed at reducing accident frequency or severity. This study is directed at the latter, specifically by seeking to understand how variables relating to the driver, road, and environment impact severity of crashes as measured by fatality.

Crash severity factor analysis today is commonly performed by constructing a mathematical model using accident data and interpreting the importance of factors towards predicting the outcome of a crash. In the modeling phase, the researcher "fits" a function which describes the relationship between the factors (independent variables) and the outcome (target variable). The types of functions historically applied range from logistic regression (Eboli et al., 2020) to more recently adopted methods such as XGBoost-SHAP (Chang et al., 2022)(Chen and Guestrin, 2016)(Lundberg and Lee, 2017). The predictiveness of the model can then be estimated using measures such as the model's AUC or RMSE values.

How much a factor contributes to the severity of crashes can then be interpreted from the parameters of the resulting model. For example, as is normally done in analyses using logistic regression, Eboli considered the coefficients and the p-values of factors to estimate their importance (Eboli et al., 2020). Chang estimates feature importance using blackbox explainers to derive SHAP values from an XGBoost model (Chang et al., 2022). Each method allows a ranking of the importance of factors and an insight into how different values within each factor impact the model's prediction of the severity of an accident's outcome.

This study follows this analysis framework and to model accident data it uses the Explainable Boosting Machine (EBM) (Nori et al., 2019), a fast implementation of the GA^2M algorithm (Lou et al., 2013) developed by Microsoft and the InterpretML community. The algorithm offers several features designed to aid interpretation of the importance of factors in the model. First, it mitigates the distorting effects of collinearity towards interpretation by restricting its boosting procedure to one feature at a time, in a round-robin fashion (Nori et al., 2019)(Wick et al., 2020). This allows the analysis to incorporate all original variables and obviates procedures to address collinearity such as finding the principal component of correlated variables or removing variables. Furthermore, the EBM implements the FAST algorithm within GA^2M, which automatically detects and ranks pairwise interaction terms with inexpensive computational cost relative to other interaction detection algorithms (Lou et al., 2013).

2 Data

As of December 2023, the FMCSA MCMIS database contains 2,967,426 records of crashes involving trucks with gross vehicle weight rating exceeding 26,000 pounds (marked as trucks with a gross vehicle weight rating of "3" in the database). Out of these, 98,745 crashes were recorded to result in at least one fatality. This study selected 11 factors associated with each crash as predictive features, with the target variable being a boolean categorical feature indicating whether a crash resulted in at least one fatality.

Table 1: Summary of MCMIS Heavy-Duty Truck Accident Data

Variables		
Column Name	Categories/Range	Completeness
TRAFFICWAY_ID	(1) Two-way, not divided (2) Two-way, divided, unprotected (3) Two-way, divided, positive barrier (4) One-way, not- divided (98) Not Reported (99) Unknown	86.9%
ACCESS_CONTROL_ID	(1) Full Control (2) Partial Access (3) No Control	70.9%
ROAD_SURFACE_CONDITION_ID	(1) Dry (2) Wet (3) Water (Standing, Moving) (4) Snow (5) Slush (6) Ice (7) Sand, Mud, Dirt Oil, or Grain (8) Other (9) Unknown	98.1%
AXLES	Range: 3 to 8 Mean: 4.46 Median: 5 Mode: 5	17.2%
GVWR (Gross Vehicle Weight Rating)	Range: 26,001 to 170,000 lbs Mean: 74,286 lbs Median: 80,000 lbs Mode: 80,000 lbs	18.1%
CARGO_BODY_TYPE_ID	(1) Bus (9 - 15 People) (2) Bus (>15 People) (3) Van/Enclosed Box (4) Cargo Tank (5) Flatbed (6) Dump (7) Concrete Mixer (8) Auto Transporter (9) Garbage/Refuse (10) Grain, Chips, Gravel (11) Pole (12) Not Applicable / No Cargo Body (13) Intermodal (14) Logging (15) Vehicle Towing Another Vehicle (98) Other	99.5%

Table 2: Summary of MCMIS Heavy-Duty Truck Accident Data (continued)

Variables		
Column Name	Categories/Range	Completeness
VEHICLE_CONFIGURATION_ID	(1) Passenger Car (2) Light Truck (3) Bus (4) Bus (5) Single Unit Truck (2 Axle, 6 Tire) (6) Single Unit Truck (3 or More Axles) (7) Truck/Trailer (8) Truck Tractor (Bobtail) (9) Tractor/Semi-Trailer (10) Tractor/Double (11) Tractor/Triples (12) Unkown Heavy Truck, Cannot Classify	86.9%
VEHICLE_HAZMAT_CLASS_ID	(1) Explosives (2) Gases-Compressed, Dissolved, or Refrigerated (3) Flammable Liquids (4) Flammable Solids - Combustible, Water Reactives (5) Oxidizing Substances - Organic Peroxides (6) Poisonous (Toxic) / Infectious Substances (7) Radioactive (8) Corrosives (9) Miscellaneous Dangerous Goods	2.2%
WEATHER_CONDITION_ID	(1) No Adverse Conditions (2) Rain (3) Sleet, Hail (4) Snow (5) Fog (6) Blowing Sand, Soil, Dirt, or Snow (7) Severe Crosswinds (8) Other (9) Unknown	98.0%
LIGHT_CONDITION_ID	(1) Daylight (2) Dark - Not Lighted (3) Dark - Lighted (4) Dark - Unknown (5) Dawn (6) Dusk (8) Other (9) Unknown	98.1%
DRIVER_CONDITION_CODE	(1) Appeared Normal (2) Had Been Drinking (3) Illegal Drug Use (4) Sick (5) Fatigue (6) Asleep (7) Medication (8) Unknown	16.2%
Target Variable - FATAL	(0) Non-Fatal (1) Fatal	

3 Methodology

3.1 Overview of EBM and GA2M

The Explainable Boosting Machine is an implementation of the GA^2M model of the format:

$$g(E[y]) = \beta_0 + \sum f_i(x_i) + \sum f_{i,j}(x_i, x_j) \tag{1}$$

Where:

1. g represents the link function which adapts to settings such as classification or regression
2. $\sum f_i(x_i)$ represents additive models each trained on one feature
3. $\sum f_{i,j}(x_i, x_j)$ represents additive models each trained on pairwise features

In the EBM, each additive model f_i and $f_{i,j}$ is a function obtained through modern machine learning techniques such as boosting or bagging (Nori et al., 2019). In practice, each additive model outputs a prediction contribution, the contributions are then summed up and inputted through the link function g which outputs the final prediction (Nori, 2022). The final models trained in this study can be categorized as binary classifiers given the boolean nature of the outcome of interest (fatal if a crash resulted in at least one fatality, non-fatal otherwise).

3.2 How the EBM ranks features

To measure the importance of features, the EBM computes the mean absolute score of each singular and pairwise feature. The mean absolute score is generally defined as:

$$Z = \frac{\sum n_i |z_i|}{\sum n_i} \tag{2}$$

Where:

1. n_i represents the number of data points in a particular category (or bin for continuous features)
2. z_i represents the contribution score the model learns for a particular category (or bin). The contribution score is in log odds for classification models.

The mean absolute score Z can be thought of as the magnitude by which a feature changes the overall prediction of the classifier (Nori, 2022). The EBM ranks singular and pairwise features based on their mean absolute scores and provides graphical representation of this ranking and of the contribution score of each category (or bin) within each feature.

3.3 Implementation

This study trained five models using stratified K-fold cross validation where K = 5 and train-test split is 80-20. Given the imbalance between positive and negative cases in the dataset (only 3.328% of cases are positive or fatal cases), stratified K-fold allows both the training and the test set to retain this ratio. Maintaining this balance allows the model to learn from enough positive cases while ensuring that the test set used to evaluate it also contains positive cases, which yields a better assessment of each of the models' AUC scores.

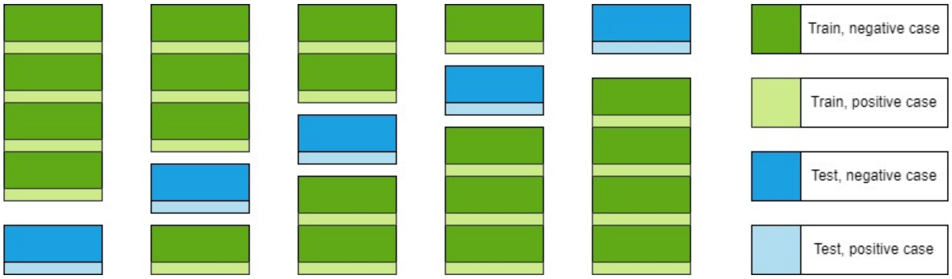


Figure 1: Stratified K-fold sampling with K = 5 (train/test proportion not drawn to reflect dataset used in this study)

3.4 Model Results

The models were trained in Python; the random seed was set to 42 to obtain the following results:

Table 3: Summary of Models

Model	AUC	Feature Importance Ranking	Mean Absolute Score (Z)
EBM 1	0.627	(1) TRAFFICWAY_ID & LIGHT_CONDITION_ID	.0657
		(2) TRAFFICWAY_ID & DRIVER_CONDITION_CODE	.0615
		(3) TRAFFICWAY_ID & ROAD_SURFACE_CONDITION_ID	.0605
		(4) ROAD_SURFACE_CONDITION_ID & LIGHT_CONDITION_ID	.0482
		(5) ACCESS_CONTROL_ID & LIGHT_CONDITION_ID	.0353
		(6) LIGHT_CONDITION_ID & DRIVER_CONDITION_CODE	.0343
		(7) AXLES & LIGHT_CONDITION_ID	.0327
		(8) CARGO_BODY_TYPE_ID & LIGHT_CONDITION_ID	.0321
		(9) WEATHER_CONDITION_ID & LIGHT_CONDITION_ID	.0294
		(10) TRAFFICWAY_ID	.0266
		(11) LIGHT_CONDITION_ID	.0236
		(12) LIGHT_CONDITION_ID & HAZMAT	.0162
		(13) ROAD_SURFACE_CONDITION_ID	.0130
		(14) ACCESS_CONTROL_ID	.0100
		(15) WEATHER_CONDITION_ID	.0099
EBM 2	0.628	(1) TRAFFICWAY_ID & LIGHT_CONDITION_ID	.0770
		(2) TRAFFICWAY_ID & ROAD_SURFACE_CONDITION_ID	.0697
		(3) TRAFFICWAY_ID & DRIVER_CONDITION_CODE	.0668
		(4) ROAD_SURFACE_CONDITION_ID & LIGHT_CONDITION_ID	.0542
		(5) ACCESS_CONTROL_ID & LIGHT_CONDITION_ID	.0338
		(6) CARGO_BODY_TYPE_ID & LIGHT_CONDITION_ID	.0333
		(7) VEHICLE_CONFIGURATION_ID & LIGHT_CONDITION_ID	.0318
		(8) LIGHT_CONDITION_ID & DRIVER_CONDITION_CODE	.0314
		(9) WEATHER_CONDITION_ID& LIGHT_CONDITION_ID	.0313
		(10) AXLES & LIGHT_CONDITION_ID	.0289
		(11) TRAFFICWAY_ID	.0272
		(12) LIGHT_CONDITION_ID	.0233
		(13) ROAD_SURFACE_CONDITION_ID	.0132
		(14) ACCESS_CONTROL_ID	.0109
		(15) WEATHER_CONDITION_ID	.0094
EBM 3	0.619	(1) TRAFFICWAY_ID & DRIVER_CONDITION_CODE	.0615
		(2) TRAFFICWAY_ID & LIGHT_CONDITION_ID	.0595
		(3) ROAD_SURFACE_CONDITION_ID & LIGHT_CONDITION_ID	.0517
		(4) ACCESS_CONTROL_ID & LIGHT_CONDITION_ID	.0353
		(5) AXLES & LIGHT_CONDITION_ID	.0346
		(6) WEATHER_CONDITION_ID& LIGHT_CONDITION_ID	.0306
		(7) TRAFFICWAY_ID	.0270
		(8) CARGO_BODY_TYPE_ID & LIGHT_CONDITION_ID	.0261
		(9) LIGHT_CONDITION_ID & DRIVER_CONDITION_CODE	.0253
		(10) LIGHT_CONDITION_ID	.0239
		(11) ROAD_SURFACE_CONDITION_ID	.0134
		(12) LIGHT_CONDITION_ID & HAZMAT	.0125
		(13) VEHICLE_CLASS_HAZMAT_ID & LIGHT_CONDITION_ID	.0119
		(14) ACCESS_CONTROL_ID	.0106
		(15) WEATHER_CONDITION_ID	.0103

Table 4: Summary of Models

Model	AUC	Feature Importance Ranking	Mean Absolute Score (Z)
EBM 4	0.617	(1) TRAFFICWAY_ID & DRIVER_CONDITION_CODE	.0558
		(2) TRAFFICWAY_ID & LIGHT_CONDITION_ID	.0530
		(3) ROAD_SURFACE_CONDITION_ID & LIGHT_CONDITION_ID	.0430
		(4) ACCESS_CONTROL_ID & LIGHT_CONDITION_ID	.0341
		(5) AXLES & LIGHT_CONDITION_ID	.0294
		(6) LIGHT_CONDITION_ID & DRIVER_CONDITION_CODE	.0292
		(7) WEATHER_CONDITION_ID & LIGHT_CONDITION_ID	.0283
		(8) CARGO_BODY_TYPE_ID & LIGHT_CONDITION_ID	.0228
		(9) TRAFFICWAY_ID	.0169
		(10) LIGHT_CONDITION_ID	.0156
		(11) LIGHT_CONDITION_ID & HAZMAT	.0122
		(12) VEHICLE_CLASS_HAZMAT_ID & LIGHT_CONDITION_ID	.0119
		(13) ROAD_SURFACE_CONDITION_ID	.0086
		(14) ACCESS_CONTROL_ID	.0074
		(15) WEATHER_CONDITION_ID	.0072
EBM 5	0.623	(1) TRAFFICWAY_ID & LIGHT_CONDITION_ID	.0714
		(2) TRAFFICWAY_ID & DRIVER_CONDITION_CODE	.0581
		(3) ROAD_SURFACE_CONDITION_ID & LIGHT_CONDITION_ID	.0572
		(4) TRAFFICWAY_ID	.0452
		(5) LIGHT_CONDITION_ID	.0409
		(6) WEATHER_CONDITION_ID& LIGHT_CONDITION_ID	.0315
		(7) ACCESS_CONTROL_ID & LIGHT_CONDITION_ID	.0304
		(8) AXLES & LIGHT_CONDITION_ID	.0292
		(9) CARGO_BODY_TYPE_ID & LIGHT_CONDITION_ID	.0283
		(10) LIGHT_CONDITION_ID & DRIVER_CONDITION_CODE	.0241
		(11) VEHICLE_CONFIGURATION_ID & LIGHT_CONDITION_ID	.0240
		(12) ROAD_SURFACE_CONDITION_ID	.0219
		(13) ACCESS_CONTROL_ID	.0164
		(14) WEATHER_CONDITION_ID	.0157
		(15) VEHICLE_CLASS_HAZMAT_ID & LIGHT_CONDITION_ID	.0134

4 Findings

4.1 Summary

In all models trained, trafficway type is the most important singular feature among the factors analyzed. On average, heavy duty truck crashes on undivided two-way lanes are 1.7 times more likely to be fatal than those occurring on two-way lanes divided by a positive barrier (see figure 2). The models also indicate that light condition impacts the severity of crashes, with results showing that crashes which occur in dark, unlighted areas are 1.8 times more likely to be fatal than those which occurred during daytime (Appendix 1). Road surface conditions also appeared to be a significant factor in severity of crashes - those which occurred in dry conditions shown to have higher likelihood of fatality than those which occurred in wet, snow, ice, or slush conditions (Appendix 2).

Reasons for the above findings can be inferred. Crashes which occur on undivided two-way lanes are likelier to be head-on collisions, which are less elastic and are generally more severe due to the greater kinetic energy dissipation relative to other types of collisions. Furthermore, drivers in dark, unlighted areas can be inferred to have lower anticipation time and, therefore, lower opportunity to transfer kinetic energy by braking prior to the collision. Finally, while wet and icy conditions are normally thought to increase the *occurrence* of crashes, they have been shown to slow overall speed of traffic flow on highways and therefore lowering the severity of crashes (Chin et al., 2004) (Shah et al., 2003) (Smith et al., 2003).

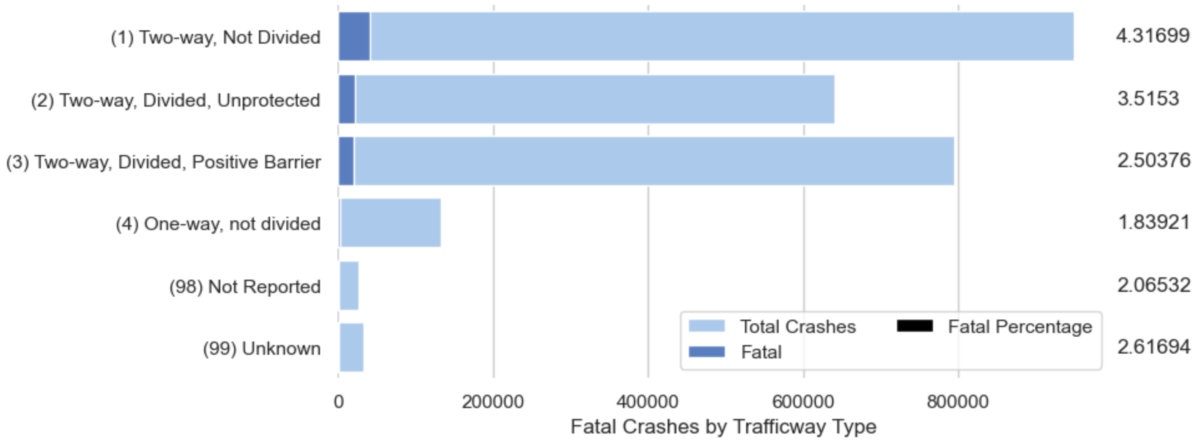


Figure 2: Fatality Rates at Different Trafficway Types

With respect to pairwise features, interaction between trafficway type and light condition appears to be most important. Crashes which occurred on undivided, two-way lanes and in dark, unlighted areas are 3.6 times more likely to be fatal compared to those which occurred in daylight and on two-way lanes divided by a positive barrier (Appendix 3). The models also rank highly the interaction between trafficway type and road surface conditions. Crashes which occured on dry surface and undivided two-way lanes are 2.2 times likelier to be fatal than those which occured on wet surface and two-way lanes with positive barrier, and over 300 times more likely to be fatal than crashes which occured in icy conditions and on two-way lanes divided by a positive barrier (Appendix 4).

The models do not rank driver condition as important as a singular feature, however when combined with trafficway type the interaction of the two variables are ranked in the top 3 in four out of five models (it is the most important feature in models 3 and 4). Crashes which occur on undivided, two-way lanes are 7.8 times more likely to be fatal when illegal drug use was reported and 2.3 times more likely to be fatal when alcohol use was reported, compared to crashes where the driver appeared normal. When drug use was reported, a crash is 1.7 times more likely to be fatal in a two-way, undivided lanes than on those divided by a positive barrier (for alcohol use, the factor is 1.15).

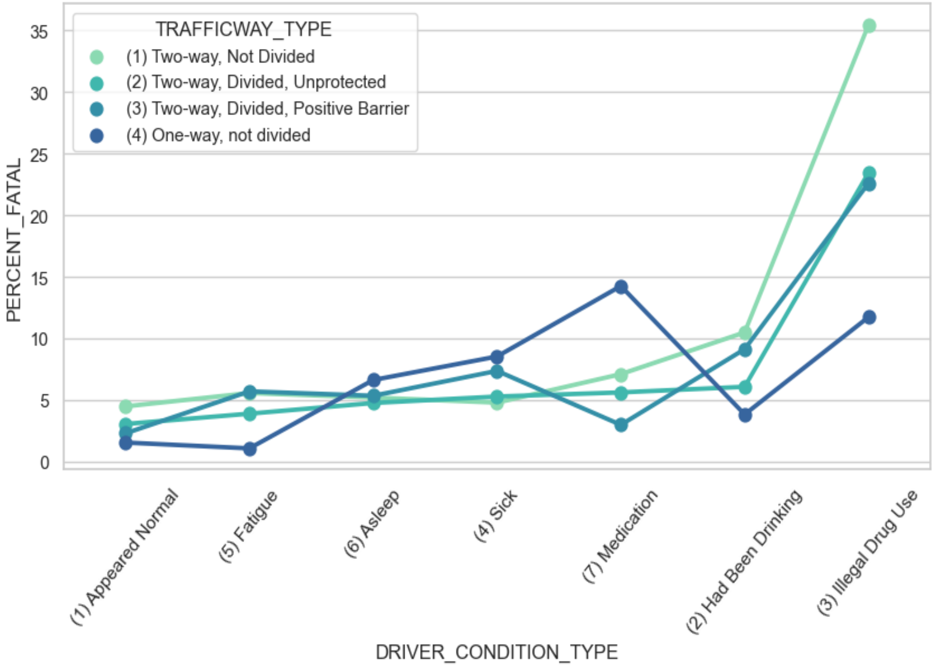


Figure 3: Interaction Plot Between Trafficway Type and Driver Condition

5 Ideas for Continuing Studies and Possible Countermeasures

To inform countermeasures, further analyses may be performed such as identifying highways with features associated with high fatality rates. For example, examination of particularly dangerous areas such as those with undivided two-lane highways and where lighting is limited during night time may be performed. The research may also look into specific locations where the crashes in this category have occurred in the past and analyze other factors impacting severity that exist within the identified areas. Countermeasures such as placing positive barriers may be performed, and effectiveness may be evaluated by examining the rates of fatality before and after the implementation of the countermeasures.

References

I. Chang, H. Park, E. Hong, J. Lee, and N. Kwon. Predicting effects of built environment on fatal pedestrian accidents at location-specific level: Application of xgboost and shap. *Accident Analysis and Prevention 166 (2022), 106545*, 2022.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. URL https://api.semanticscholar.org/CorpusID:4650265.

S. M. Chin, O. Franzese, D. L. Greene, H. H. L., and R. C. Gibson. Temporary losses of highway capacity and impacts on performance: Phase 2. 11 2004. doi: 10.2172/885576. URL https://www.osti.gov/biblio/885576.

L. Eboli, C. Forciniti, and G. Mazzulla. Factors influencing accident severity: an analysis by road accident type. *Transportation Research Procedia 47 (2020), 449-456*, 2020.

Y. Lou, R. Caruana, J. Gehrke, and G. Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Neural Information Processing Systems*, 2017. URL https://api.semanticscholar.org/CorpusID:21889700.

H. Nori. Mean absolute score : Overall importance, issue 337, 2022. URL https://github.com/interpretml/interpret/issues/337.

H. Nori, S. Jenkins, P. Koch, and R. Caruana. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019.

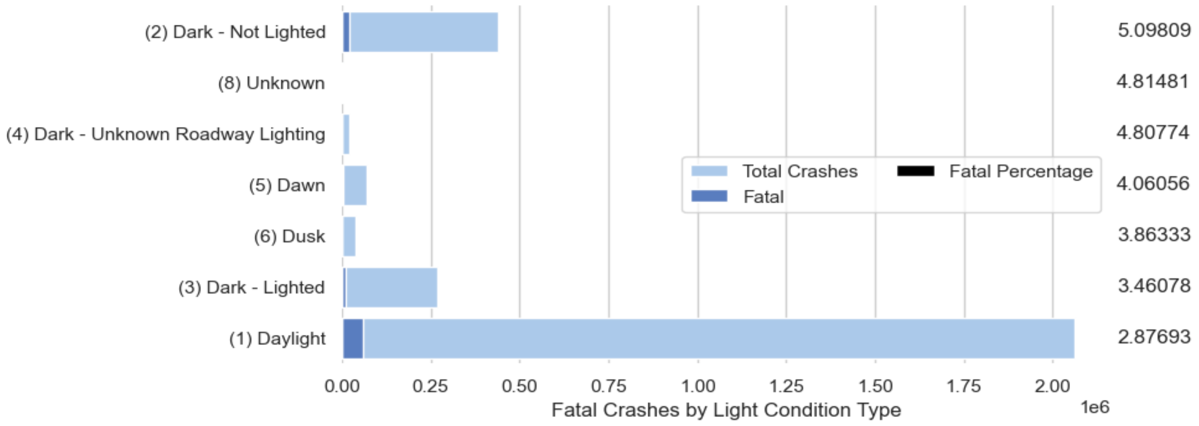
V. P. Shah, A. D. Stern, L. Goodwin, and P. A. Pisano. Analysis of weather impacts on traffic flow in metropolitan washington, dc. 2003. URL https://api.semanticscholar.org/CorpusID:126833145.

B. Smith, K. Byrne, R. Copperman, S. Hennessy, and N. Goodall. An investigation into the impact of rainfall on freeway traffic flow. 2003. doi: https://doi.org/10.31224/osf.io/9xnzc.

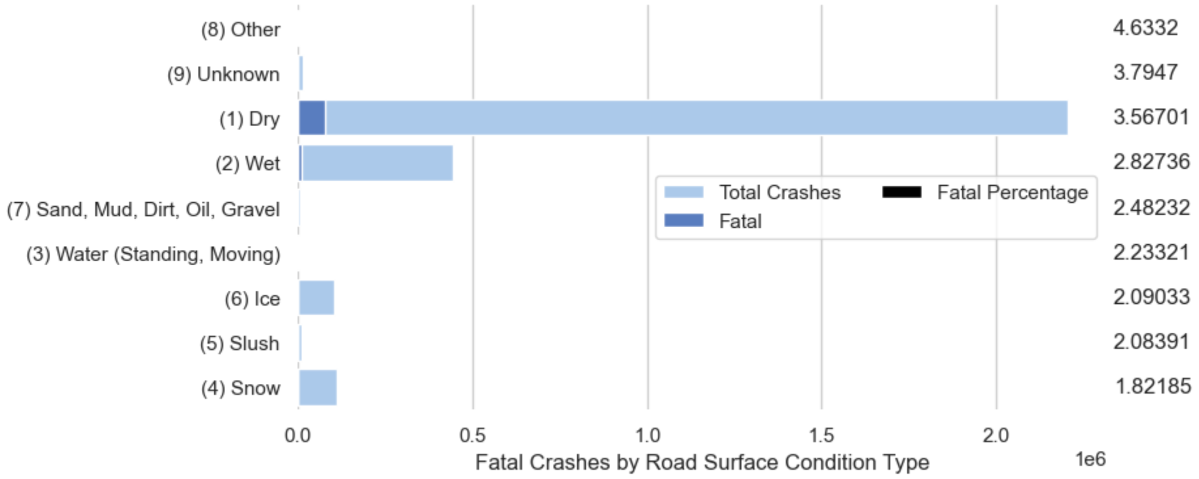
F. Wick, U. Kerzel, and M. Feindt. Cyclic boosting - an explainable supervised machine learning algorithm. *CoRR*, abs/2002.03425, 2020. URL https://arxiv.org/abs/2002.03425.

6 Appendix

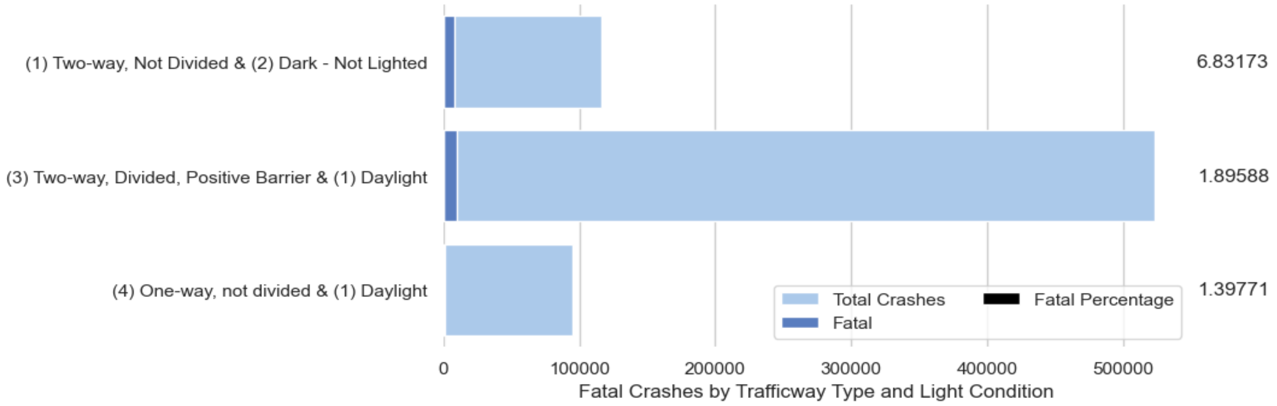
6.1 Appendix 1 - Fatality Rates at Different Light Conditions



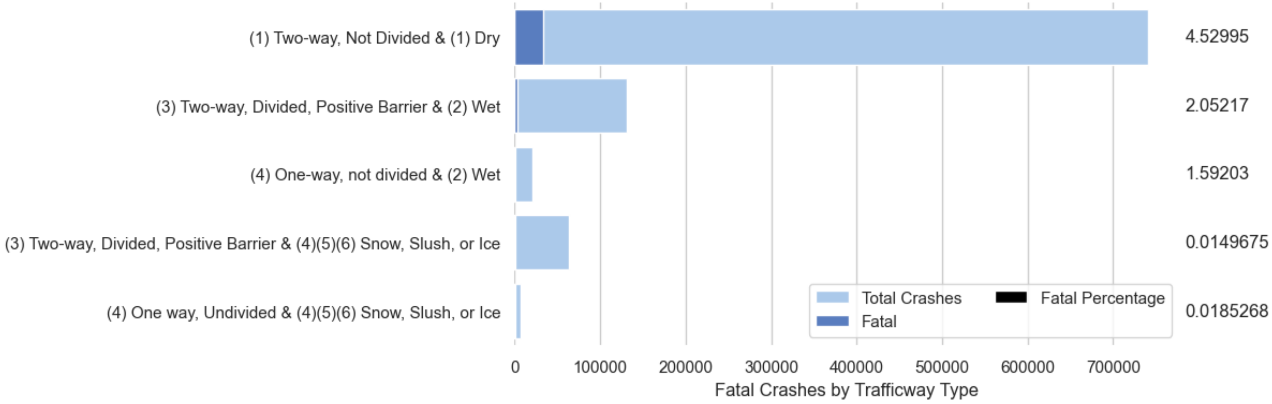
6.2 Appendix 2 - Fatality Rates at Different Road Surface Conditions



6.3 Appendix 3 - Interaction between Trafficway Types and Light Conditions



6.4 Appendix 4 - Interaction between Trafficway Types and Surface Conditions


(1) Two-way, Not Divided & (1) Dry
(3) Two-way, Divided, Positive Barrier & (2) Wet
(4) One-way, not divided & (2) Wet
(3) Two-way, Divided, Positive Barrier & (4)(5)(6) Snow, Slush, or Ice
(4) One way, Undivided & (4)(5)(6) Snow, Slush, or Ice
Total Crashes
Fatal
Fatal Percentage
Fatal Crashes by Trafficway Type