Drunk Driver Involved Fatal Car Crash Prediction

Nick House

Introduction to Artificial Intelligence

May, 2020

Contents

List of Figures

List of Tables

1 Introduction

Fatal traffic crashes accounted for 36,560 fatalities in the United States and United States Territories in the year 2018, according to the National Center for Statistics and Analysis (NCSA) of the National Highway Traffic Safety Administration (NHTSA). The NCSA is responsible for providing a wide range of analytical and statistical support to the NHTSA and the highway safety community at large. There is a vast amount of data collected every year and made available by the NCSA to the public. The NCSA available data goes as far back as the 1970's. With the magnitude of data and data elements collected every year, there is many options for making predictions. I chose to see if I could train a model to predict whether there was a drinking driver involved in a traffic fatality given 28 input features. Out of 36,560 traffic fatalities 9,308 of them involved a drinking driver in the year 2018. 24.9% of all fatal crashes involved a drinking driver, accounting for 25.5% of all crash fatalities. I would like to potentially identify situations or circumstances that could be addressed to reduce fatal drunk driving crashes. I believe features such as day of the week , time of the day, type of road and type of intersections would have a significant indication on fatalities and I suspect these would have some correlation to if there was a drinking driver involved.

2 Dataset

The dataset I chose was from the Fatality Analysis Reporting System (FARS), every row of data represents a crash that resulted in at least one fatality in the year 2018. The data was populated by traffic fatality reports filed from all 50 states and 3 US territories (District of Columbia, Virgin Islands and Puerto Rico).

2.1 Dataset Description

The dataset originally contained 33,655 rows and 51 columns with 1.73 million data points, majority being categorical data. For this project I significantly reduced the dataset to 28 input features with a single output feature.
(See Appendix A for detailed feature descriptions)

2.2 Data Processing

Many columns were removed due to them not pertaining to the nature of the crash itself. This would include features like case numbers and times of actions completed after the crash, such as emergency response times to the scene of the crash. For crash reporting times I categorized them into one of five blocks representing different sections of a 24 hour period. For the remaining categorical data I cast all input values to be sequential, alleviating any categorical outlying values.

2.3 Data Visualization

(For individual feature histograms see Appendix B)

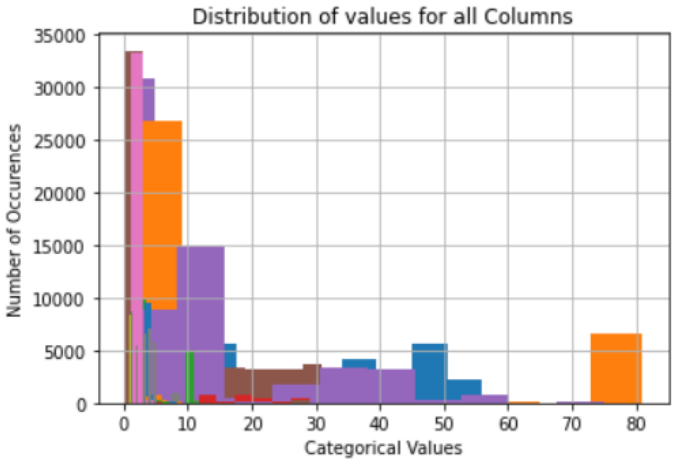


Figure 1: Feature Distribution

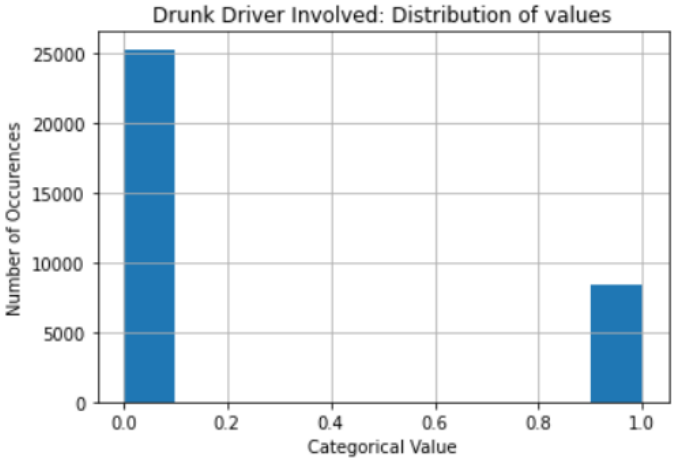


Figure 2: Output Distribution

Feature	Mode	Minimum	Maximum	Median	Mean
STATE	48	1	56	26	27.2%
PVH	0	0	20	0	.04%
PERNOTMVIT	0	0	10	0	.24%
PERMVIT	1	0	47	2	2.24%
PERSONS	1	0	47	2	2.25%
DAY	3	1	31	15	15.52%
MONTH	10	1	12	7	6.65%
DAYofWEEK	7	1	7	4	4.11%
NHS	0	0	3	0	.44%
Rur Urb	2	1	3	2	1.58%
FUNC SYS	3	1	8	4	3.89%
RD OWNER	1	1	81	1	17.41%
ROUTE	3	1	9	3	3.83%
SP JUR	0	0	9	0	.03%
HARM EV	12	1	75	12	17.67%
MAN COLL	0	0	9	0	1.00%
RELJCT1	0	0	2	0	.05%
RELJCT2	1	1	14	1	1.94%
TYP INT	1	1	9	1	1.38%
WRK ZONE	0	0	4	0	.o4%
REL ROAD	1	1	12	1	2.17%
LGT COND	1	1	9	2	1.92%
WEATHER1	1	0	12	1	2.45%
CF1	0	0	29	0	1.48%
CF2	0	0	29	0	.40%
CF3	0	0	29	0	.29%
FATALS	1	1	20	1	1.09%
MinutesOfDay	1	1	5	3	2.87%
Drunk Driver	0	0	1	0	.25%

Table 1: Feature Statistics

3 Data Processing

3.1 Training and Validation Sets

Train Test Split was utilized to split data. Data was seeded and shuffled, then ran through several iterations of different size splits to find an ideal training/validation split ratio. A 15%/85% split showed a small advantage in accuracy over the others tested.

Test Split Ratio	Validation Accuracy
10/90	78.82%
15/85	79.08%
20/80	79.04%
30/70	78.57%

Table 2: Train Test Split

3.2 Data Normalization

Because categorical data is being utilized, a numerical value representing a data type has no direct numerical relation to another value representing a different datatype. Thus, typical data normalization techniques showed little to no learning during model training. So, in addition to ordering all feature categories into sequential values, One hot encoding was utilized and showed the most significant results in training models for this dataset. This reshaped the 28 input feature columns into 404 binary input columns. Each input feature's categorical value was transformed into an individual binary column. This normalized all categorical values to either a one or zero.

4 Modelling

4.1 Building a Model

The model was built and tested one piece at a time, for each iteration its behavior in relation to Validation Accuracy was considered and documented. The number of Epochs ran was in consideration of best outcome in terms of Validation Accuracy. Given the results, marginal improvement was seen between layers, further evaluation will have to be considered looking at their Precision, Recall and F1 scores.

Number of Hidden Layers	Validation Accuracy	Epochs
No Hidden Layer	79.10%	500
One Layer	79.01%	300
Two Layers	78.89%	150
Three Layers	79.18%	270
Four Layers	78.67%	300
Five Layers	78.83%	350

Table 3: Neural Network Layer Testing

4.2 Testing Different Activation

Testing various activation with logistic regression, each ran for 350 epochs.

Activation Layer	Activation Output Layer	Validation Accuracy
Sigmoid	Relu	78.43%
Relu	Relu	76.59%
Relu	Sigmoid	77.50%
Linear	Sigmoid	78.97%
Linear	Linear	69.97%
Sigmoid	Linear	78.06%
Sigmoid	Sigmoid	79.07%

Table 4: Testing Differing Activation

4.3 Learning Curves

The following learning curves are the validation loss and accuracy curves for a single hidden layer neural network.

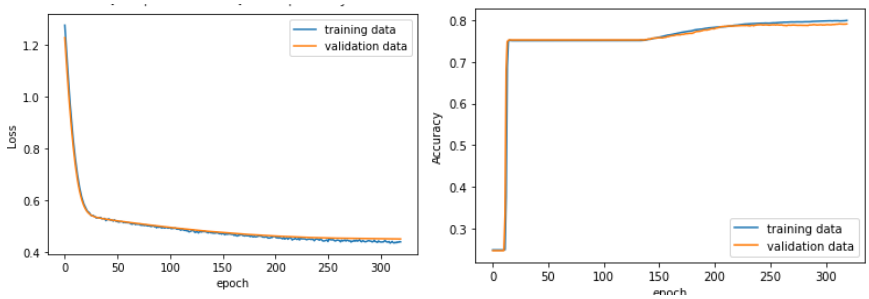


Figure 3: Learning Curve Single Hidden Layer

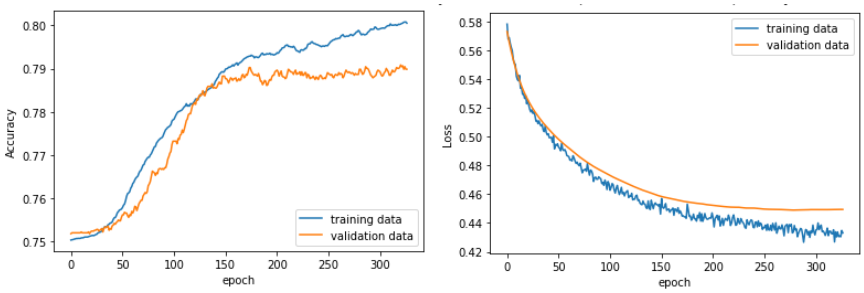


Figure 4: Learning Curve Logistic Regression

4.4 Overfitting Data

There are a number of ways to overfit data, but given a fairly robust model it can still be overfit by simply increasing the number of epochs. Below are two separate learning curves attempting to overfit both models (without output as input and with output as input) by increasing epochs. With the model not using output as input it can be observed that the loss of the training data continues to decrease as epochs increase. However, at a point loss of the validation data actually begins to increase. Conversely, with this same model run with outputs used as inputs both training and validation data quickly approached 100% and stayed.

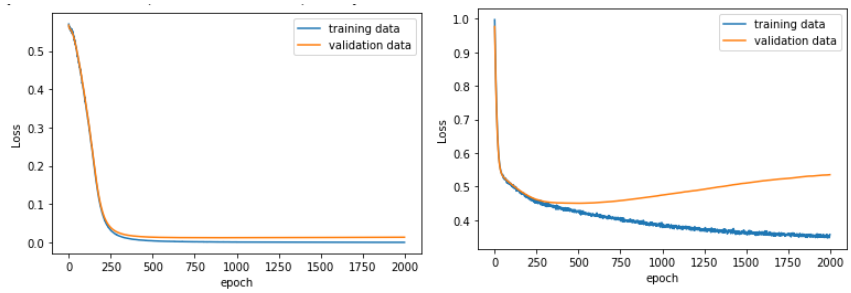


Figure 5: Overfitting Model with/without Output as Input

4.5 Overfitting Data using Output as Input

Overfitting data using output as input can be done by how training and validation data is split. By increasing validation data to 99% overfitting can be seen rather quickly.

4.6 Evaluating Model

Given the data set is predicting binary classification data and the Validation accuracy for single hidden layer neural network and logistic regression are com-

parable. Observing Recall, Precision and F1 of both models will be helpful.

Model Type	Precision	Recall	F1
Neural Network	61.29%	41.68%	.50
Linear Regression	61.53%	39.28%	.48

Table 5: Model Performance

5 Feature Importance and Evaluation

Methodology used for feature importance, evaluations and removal was done in two parts. First part was run inside a loop to test each input feature alone and measure its performance. These values were stored in an array and sorted (least significant(LS) to highest). Second part automated a loop to remove the least significant feature columns one at a time, one hot encode, train test split and run the model until the final two columns remain. Through this evaluation there showed no significant effect to the overall performance of the model. However, slight improvements were found as well as slight deterioration of the model's performance.

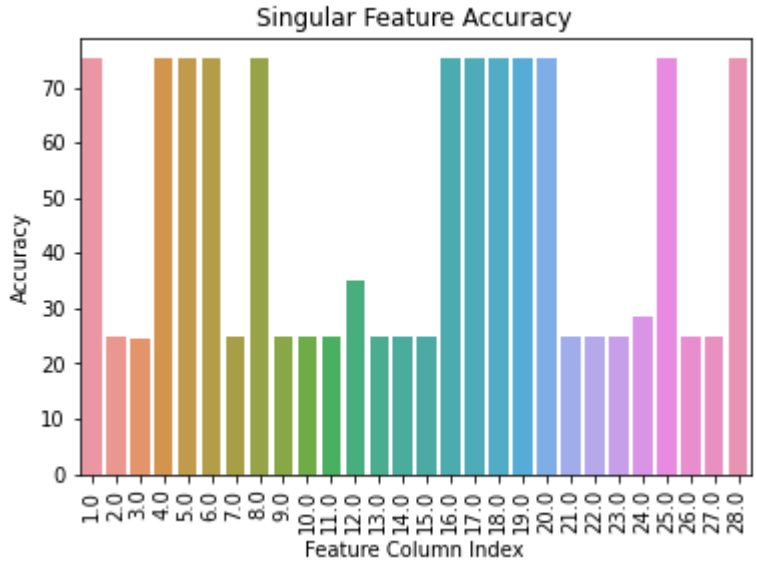


Figure 6: Single Feature Accuracy

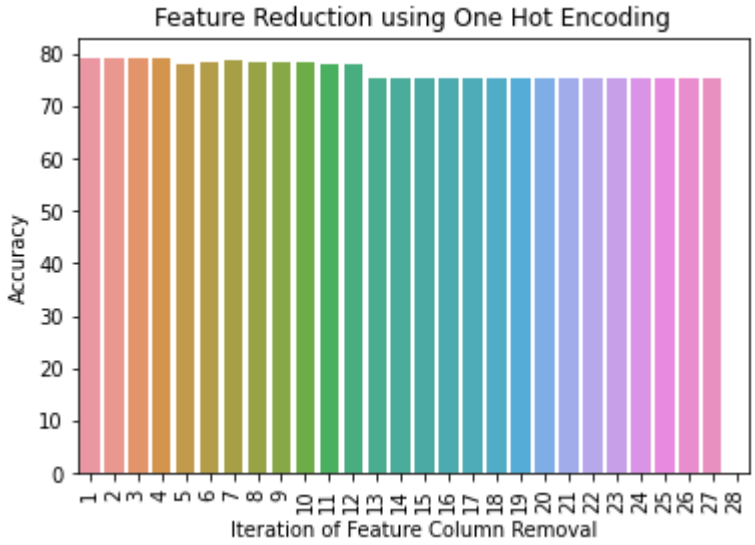


Figure 7: Model Performance by Number of Least Significant Features Removed

5.1 Performance of Least Significant Features

From testing and evaluation it can seen (below) there is a degree of improvement by removing input layers until accuracy begins to degrade. There is slight accuracy improvement by removing up to the first 4 columns, after which, accuracy begins its decline back to a ratio on par with the binary outputs value distribution (3:1).

Number of Features Removed	Accuracy	Precision	Recall	F1-Score
0	78.97%	61.22%	41.04%	.49
1	79.18%	62.24%	40.48%	.49
2	79.12%	61.98%	40.56%	.49
3	79.08%	62.16%	39.68%	.48
4	79.20%	62.47%	40.88%	.49
5	78.39%	60.67%	36.16%	.45

Table 6: Model Performance: First 5 LS Features Removed

5.2 Features Removed

A list of the inputs removed that resulted in the best accuracy performance.

Column	Feature Name	Feature Description
3	PERNOTMVIT	Number of persons not in transport
14	SP_JUR	Special Jurisdiction or not
22	LGT_COND	Light Condition if Reported
15	HARM_EV	Type of first injury.

Table 7: feature columns removed to achieve best performance

6 Cross-Validation

Cross validation was performed on the model using k-folding, folded five times. No single grouping performed noticeably better than another, which suggests the data is well distributed and the model would likely perform well given similar feature type data from another year.

Split	Accuracy
1	77.61%
2	77.91%
3	79.27%
4	78.64%
5	78.58%
Mean	78.40%
Max	79.27%
Min	77.61%
SD	0.65

Table 8: Cross Validation Statistics

7 Challenges

Challenging aspects of this project tended to be centered around data preparation and shaping the data to allow a model to learn. In the early stages of shaping and testing my data a quick climb to 75% accuracy was often seen, followed by a accuracy plateau. This quick climb gave a false sense of legitimate learning taking place. By review of the output's binary distribution 75.1%(F):24.9%(T) it was assessed that the model was likely quickly learning that a false value was correct more so than not, resulting in the model's prediction accuracy being almost identical to the percentage of false values found in the output data. Learning did not improve past 75% with any type of model changes or techniques tried until addressing the dataset itself. Ordering categorical values sequentially, adding time of crash into categories and using one hot encoding showed immediate results above all other methods tried. These methods allowed the model to continue learning beyond 75% accuracy around the 150 epoch mark.

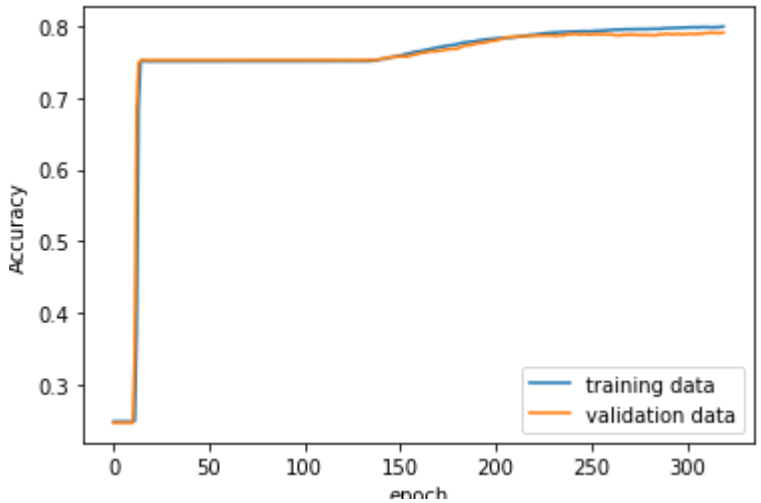


Figure 8: Accuracy can be Misleading - Learning Curve

8 Future Improvements

In the future I would like to run similar types of data from the NCSA, then validate with data from previous years. With a better understanding of how best to feature reduce and tweak model settings I believe time best served would be on really nitpicking the data sets themselves. Categorizing additional input features that may not have been included in this testing might show positive results. Utilizing an automation loop to rank features and data changes could be a tremendous help to find quality feature additions and categorizing techniques.

9 Conclusion

During this project a neural network model and logistic model showed the best results predicting whether a fatal accident involved a drinking driver. Different numbers of hidden layers, types of activation, types of optimization, feature removals, training/validation split ratios and data normalization and shaping took place. This particular project showed the best results using sigmoid activation, Adam optimization, test/validation split of 15:85, one hot encoding and either logistic regression or a single hidden layer neural network. This combination or parameters showed the best and most consistent results, it also proved to be quite robust. This model was able to interpret all input features and predict whether the crash involved a drinking driver or not with 79% accuracy. There is a vast amount of additional corresponding data available for this data set, in the future I plan to use this model as a basis to test additional data sets that will be shaped to include additional input features including non-fatal car crashes.

10 References

[1] Office of Data Acquisition, State Data Reporting Systems Division (2019). Fatality Analysis Reporting System (FARS): Analytical User's Manual, 1975-2018. Washington, DC: National Center for Statistics and Analysis National Highway Traffic Safety Administration

11 Coding and Data Links

Colab: Data Analysis

Colab: Model and Feature Removal

Link to Dataset NHTSA: Dataset

Appendices

A Data Feature Information

A.1 STATE

State in which the crash occurred
There are 56 listed categories for state. The States and US Territories are listed and numbered alphabetically (1-56). There are a combined total of 53 options for states and territories, three previously documented territories have been removed by FARS. The category values remained consistent with previously published FARS Analytical User's Manuals. The three included territories are District of Columbia, Virgin Islands and Puerto Rico.

A.2 Person Type

This data element describes the role of this person involved in the crash.

1 Driver of a Motor Vehicle In-Transport
2 Passenger of a Motor Vehicle In-Transport
3 Occupant of a Motor Vehicle Not In-Transport
4 Occupant of a Non-Motor Vehicle Transport Device
5 Pedestrian
6 Bicyclist
7 Other Cyclist
8 Person on Personal Conveyances
9 Unknown Occupant Type in a Motor Vehicle In-Transport
10 Persons In/On Buildings
19 Unknown Type of Non-Motorist

A.3 PVH

Number of Parked Working Vehicles Involved
0-999 Number of Parked/Working Vehicles in the crash

A.4 PERNOTMVIT: Number of Persons Not in Motor Vehicles in Transport

All Person records linked to the crash are used. This data element is derived as the count of all persons in the crash where "Person Type" is in (3, 4, 5, 6, 7, 8, 10 or 19).

A.5 PERMVIT

Number of Persons in Motor Vehicles in Transport
All Person records linked to the crash are used. This data element is derived as the count of all persons in the crash where “Person Type” is in (1, 2 or 9).

A.6 PERSONS

This is the count of all persons of ”Person Type” (1, 2, 3 or 9)

A.7 DAY

This records the day of the month which the crash occurred

A.8 MONTH

This records the month in which the crash occurred Jan(1)...Dec(12)

A.9 DayofWeek

This records the day of the week which the crash occured Sun(1)...Sat(7)

A.10 NHS

This data element identifies whether this crash occurred on a trafficway that is part of the National Highway System.

0 This Section is Not on the National Highway System
1 This Section is on the National Highway System
3 Unknown

A.11 Rur Urb

This identifies the classification of the segment of the traffic way on which the crash occurred based on boundaries of small urban and urbanized areas.

1 Rural
2 Urban
3 Unknown

A.12 FUNC SYS

This data element identifies the functional classification of the segment of the trafficway on which the crash occurred.

1 Interstate

2 Principal Arterial – Other Freeways and Expressways
3 Principal Arterial – Other
4 Minor Arterial
5 Major Collector
6 Minor Collector
7 Local
9 Unknown

A.13 RD OWNER

This data element identifies the entity that has legal ownership of the segment of the trafficway on which the crash occurred.

1 State Highway Agency
2 County Highway Agency
3 Town or Township Highway Agency
4 City or Municipal Highway Agency
11 State Park, Forest or Reservation Agency
12 Local Park, Forest or Reservation Agency
21 Other State Agency
25 Other Local Agency
26 Private (other than Railroad)
27 Railroad
31 State Toll Road
32 Local Toll Authority
40 Other Public Instrumentality (i.e., Airport)
50 Indian Tribe Nation
60 Other Federal Agency
62 Bureau of Indian Affairs
63 Bureau of Fish and Wildlife
64 U.S. Forest Service
66 National Park Service
67 Tennessee Valley Authority
68 Bureau of Land Management
69 Bureau of Reclamation
70 Corps of Engineers
72 Air Force
74 Navy/Marines
80 Army
81 Unknown

A.14 ROUTE

his data element identifies the route signing of the trafficway on which the crash occurred.

1 Interstate
2 U.S. Highway
3 State Highway
4 County Road
5 Local Street – Township
6 Local Street – Municipality
7 Local Street – Frontage Road
8 Other
9 Unknown

A.15 SP JUR

This data element identifies if the location on the trafficway where the crash occurred qualifies as a Special Jurisdiction even though it may be patrolled by state, county or local police (e.g., all State highways running through Indian reservations are under the jurisdiction of the Indian reservation).
0 No Special Jurisdiction (Includes National Forests)

1 National Park Service
2 Military
3 Indian Reservation
4 College/University Campus
5 Other Federal Properties
8 Other
9 Unknown

A.16 HARM EV

This data element describes the first injury or damage producing event of the crash.

There is over 50 listed damage producing events for this feature category. For detailed information on each event, see Fatality Analysis Reporting System (FARS) Analytical User's Manual, 1975-2018 (listed in Reference)

A.17 MAN COLL

This data element describes the orientation of two motor vehicles in-transport when they are involved in the "First Harmful Event" of a collision crash. If the

"First Harmful Event" is not a collision between two motor vehicles in-transport it is classified as such.

0 Not Collision with Motor Vehicle in Transport
1 Front-to-Rear
2 Front-to-Front
3 Angle
4 Sideswipe – Same Direction
5 Sideswipe – Opposite Direction
6 Rear-to-Side
7 Rear-to-Rear
8 Other (End-Swipes and Others)
9 Not Reported

A.18 RELJCT1

This data element identifies the crash's location with respect to presence in an interchange area. The coding of this data element is done in two sub-fields and is based on the location of the "First Harmful Event" of the crash.

0 No
1 Yes
2 Not Reported

A.19 RELJCT2

This data element identifies the crash's location with respect to presence in or proximity to components typically in junction or interchange areas. The coding of this data element is done in two sub-fields and is based on the location of the "First Harmful Event" of the crash.

1 Non-Junction
2 Intersection
3 Intersection Related
4 Driveway Access
5 Entrance/Exit Ramp Related
6 Railway Grade Crossing
7 Crossover Related
8 Driveway Access Related
9 Shared-Use Path Crossing
10 Acceleration/Deceleration Lane
11 Through Roadway
12 Other Location Within Interchange Area
13 Entrance/Exit Ramp

14 Not Reported

A.20 TYP INT

This data element identifies and allows separation of various intersection types.

1 Not an Intersection
2 Four-Way Intersection
3 T-Intersection
4 Y-Intersection
5 Traffic Circle
6 Roundabout
7 Five-Point, or More
8 L-Intersection
9 Not Reported

A.21 WRK ZONE

This data element identifies a motor vehicle traffic crash in which the first harmful event occurs within the boundaries of a work zone or on an approach to or exit from a work zone, resulting from an activity, behavior, or control related to the movement of the traffic units through the work zone.

0 None
1 Construction
2 Maintenance
3 Utility
4 Work Zone, Type Unknown

A.22 REL ROAD

This data element identifies the location of the crash as it relates to its position within or outside the trafficway based on the "First Harmful Event."

1 On Roadway
2 On Shoulder
3 On Median
4 On Roadside
5 Outside Trafficway
6 Off Roadway – Location Unknown
7 In Parking Lane/Zone
8 Gore
10 Separator

11 Continuous Left-Turn Lane
12 Not Reported

A.23 LGT CON

This data element records the type/level of light that existed at the time of the crash as indicated in the case material.

1 Daylight
2 Dark – Not Lighted
3 Dark – Lighted
4 Dawn
5 Dusk
6 Dark – Unknown Lighting
7 Other
8 Not Reported
9 Reported as Unknown

A.24 WEATHER1

This data element records the prevailing atmospheric conditions that existed at the time of the crash as indicated in the case material.

0 No Additional Atmospheric Conditions
1 Clear
2 Rain
3 Sleet, Hail
4 Snow
5 Fog, Smog, Smoke
6 Severe Crosswinds
7 Blowing Sand, Soil, Dirt
8 Other
10 Cloudy
11 Blowing Snow
12 Freezing Rain or Drizzle

A.25 CF1, CF2, CF3

This data element records factors related to the crash expressed by the investigating officer.

0 None
1 Inadequate Warning of Exits, Lanes Narrowing, Traffic Controls etc.

2 Shoulder Related
3 Other Maintenance or Construction-Created Condition
4 No or Obscured Pavement Marking
5 Surface Under Water
6 Inadequate Construction or Poor Design of Roadway, Bridge, etc.
7 Surface Washed Out (Caved in, Road Slippage)
12 Distracted Driver of a Non-Contact Vehicle
13 Aggressive Driving/Road Rage by Non-Contact Vehicle Driver
14 Motor Vehicle Struck By Falling Cargo or Something That Came Loose From or Something That Was Set in Motion By a Vehicle
15 Non-Occupant Struck By Falling Cargo, or Something Came Loose from or Something That Was Set In Motion By A Vehicle
16 Non-Occupant Struck Vehicle
17 Vehicle Set In Motion By Non-Driver
18 Date of Crash and Date of EMS Notification Were Not Same Day
19 Recent Previous Crash Scene Nearby
20 Police-Pursuit-Involved
21 Within Designated School Zone
22 Speed Limit Is a Statutory Limit as Recorded or Was Determined as This State's "Basic Rule"
23 Indication of a Stalled/Disabled Vehicle
24 Unstabilized Situation Began and All Harmful Events Occurred Off of the Roadway
25 Toll Booth/Plaza Related
26 Backup Due to Prior Non-Recurring Incident
27 Backup Due to Prior Crash
28 Backup Due to Regular Congestion
29 Reported as Unknown

A.26 FATALS

This data element records the number of fatally injured persons in the crash.

1-99 Number of Fatalities that Occurred in the Crash

A.27 Minutes of Day

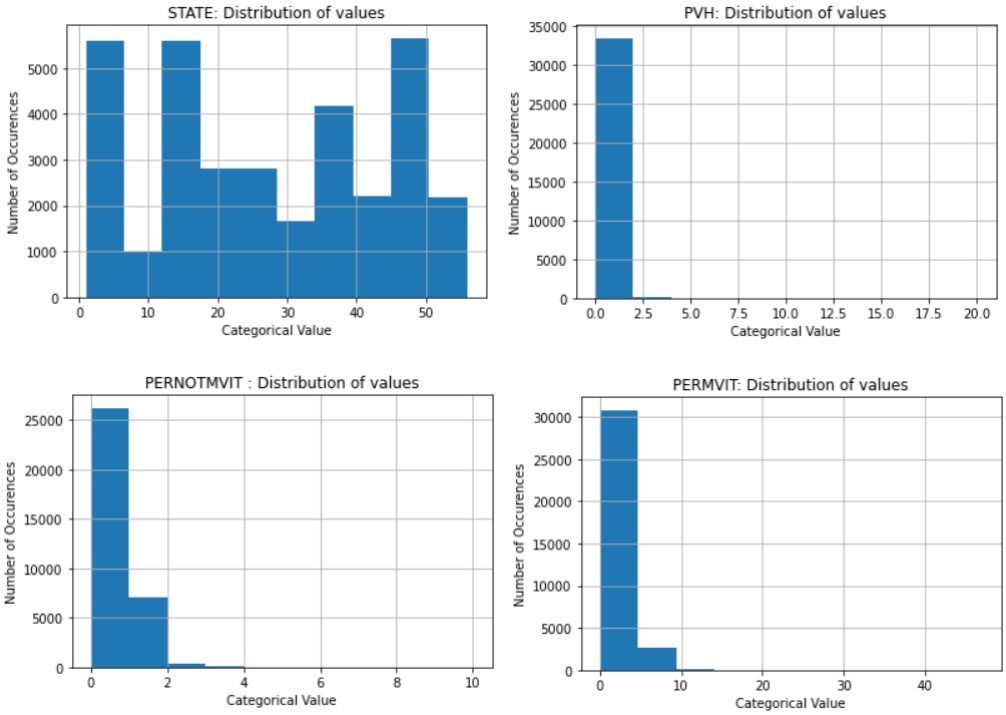
This data element records which time block of the day the crash occurred. 1 Time of Day: 0000-0500
2 Time of Day: 0500-1000
3 Time of Day: 1000-1400
4 Time of Day: 1400-1900
5 Time of Day: 1900-2400

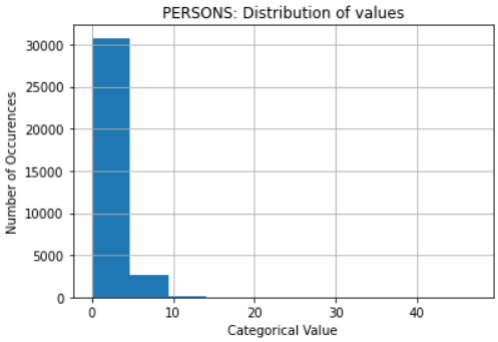
A.28 Drunk Driver Involved

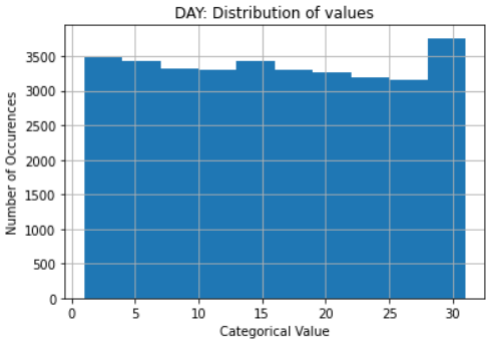
This data element records whether a drunk driver was involved in the crash or not.
0 No Drunk Drivers Involved
1 Yes Drunk Drivers Involved

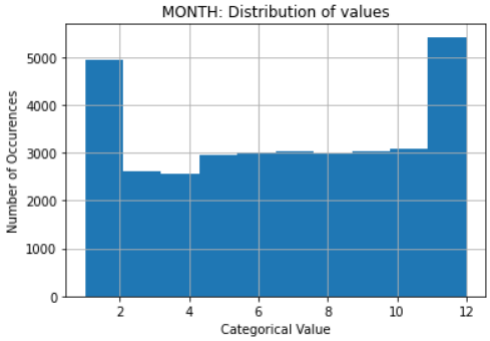
B Data Feature Distribution

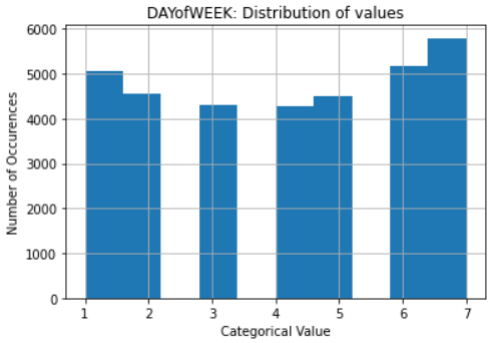
B.1 Histograms

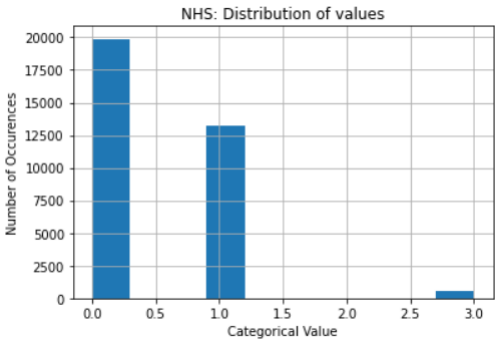


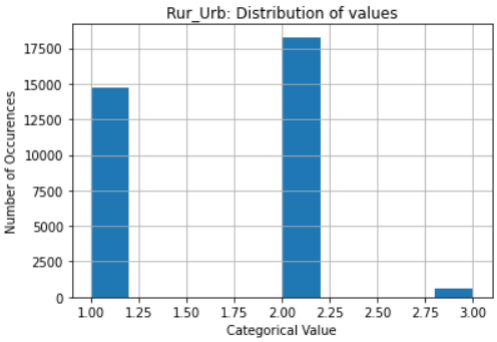

PERSONS: Distribution of values
Number of Occurences
Categorical Value

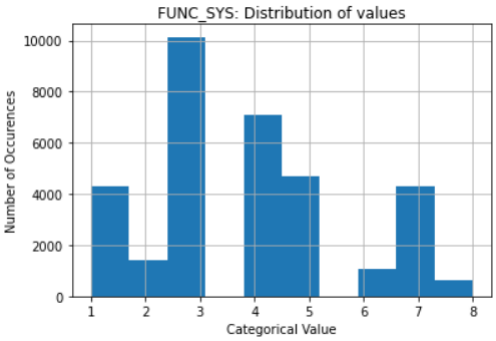

DAY: Distribution of values
Number of Occurences
Categorical Value

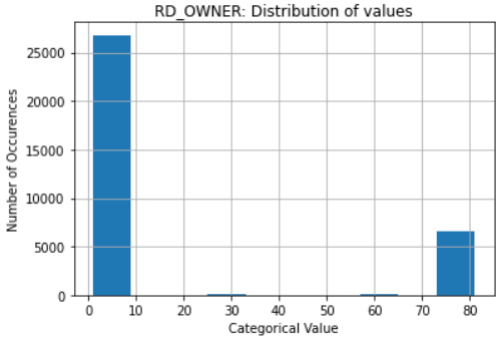

MONTH: Distribution of values
Number of Occurences
Categorical Value

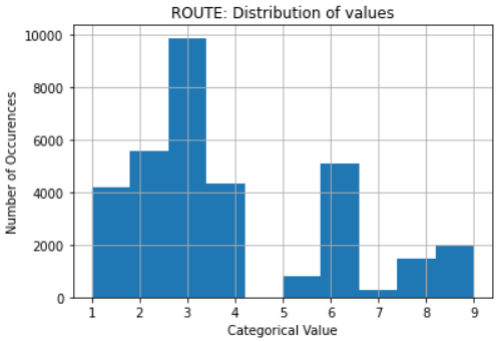

DAYofWEEK: Distribution of values
Number of Occurences
Categorical Value

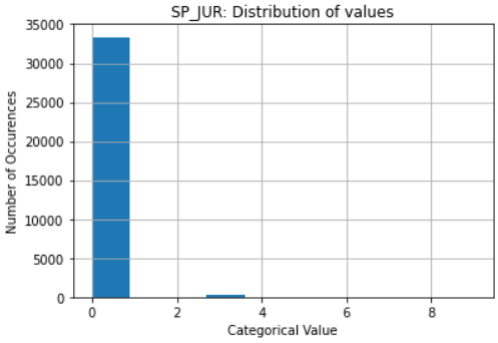

NHS: Distribution of values
Number of Occurences
Categorical Value

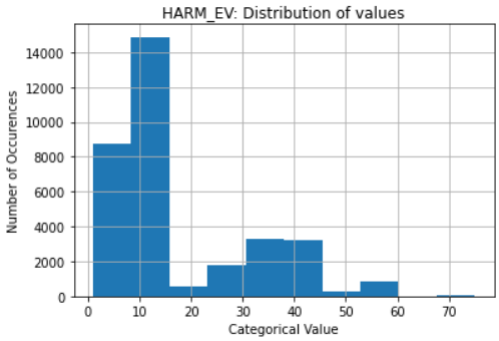

Rur_Urb: Distribution of values
Number of Occurences
Categorical Value

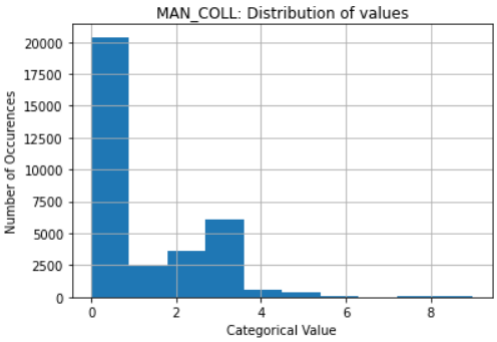

FUNC_SYS: Distribution of values
Number of Occurences
Categorical Value

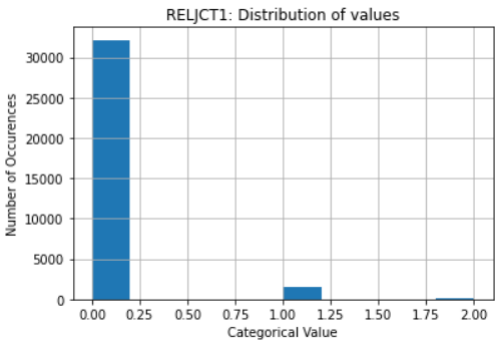

RD_OWNER: Distribution of values
Number of Occurences
Categorical Value

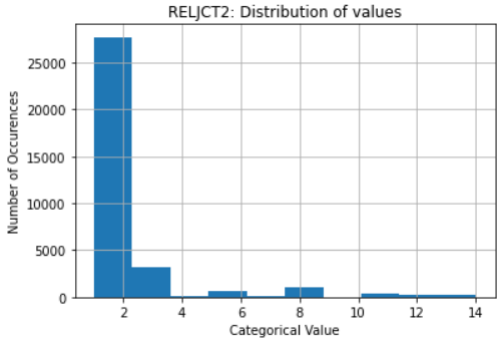

ROUTE: Distribution of values
Number of Occurences
Categorical Value

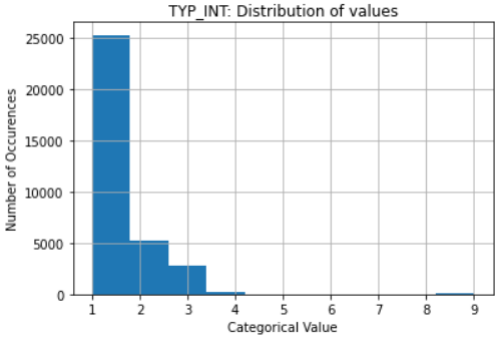

SP_JUR: Distribution of values
Number of Occurences
Categorical Value

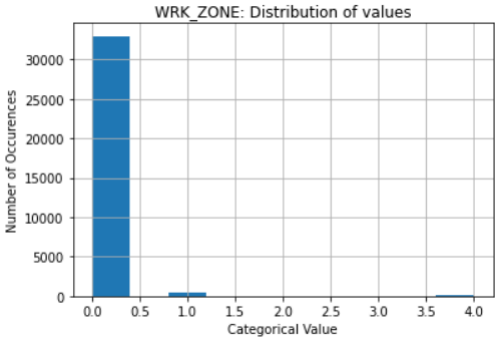

HARM_EV: Distribution of values
Number of Occurences
Categorical Value

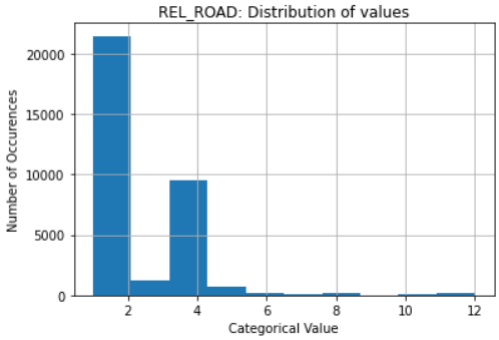

MAN_COLL: Distribution of values
Number of Occurences
Categorical Value

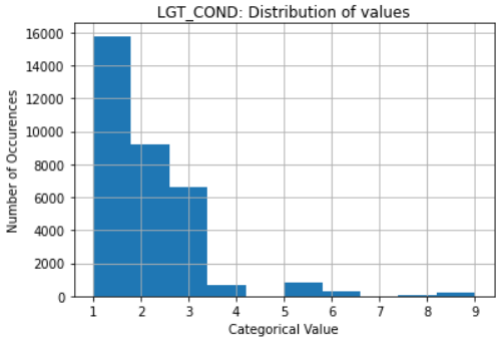

RELJCT1: Distribution of values
Number of Occurences
Categorical Value

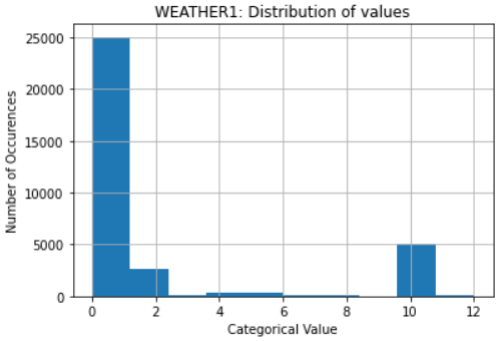

RELJCT2: Distribution of values
Number of Occurences
Categorical Value

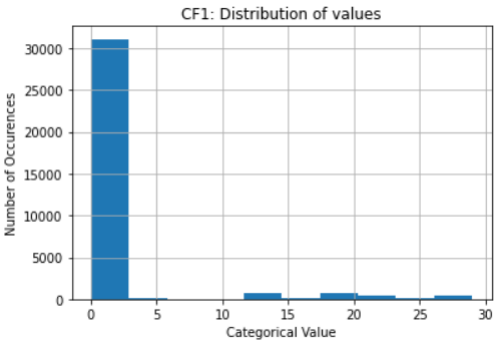

TYP_INT: Distribution of values
Number of Occurences
Categorical Value

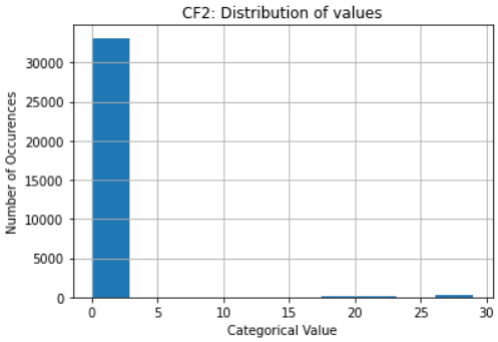

WRK_ZONE: Distribution of values
Number of Occurences
Categorical Value

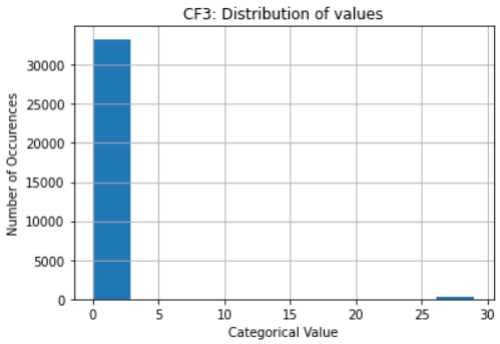

REL_ROAD: Distribution of values
Number of Occurences
Categorical Value

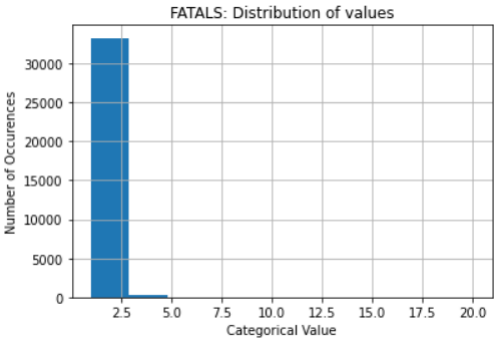

LGT_COND: Distribution of values
Number of Occurences
Categorical Value

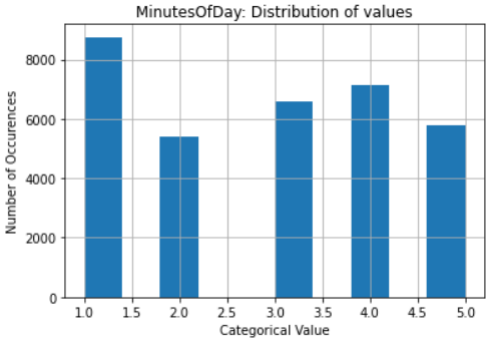

WEATHER1: Distribution of values
Number of Occurences
Categorical Value

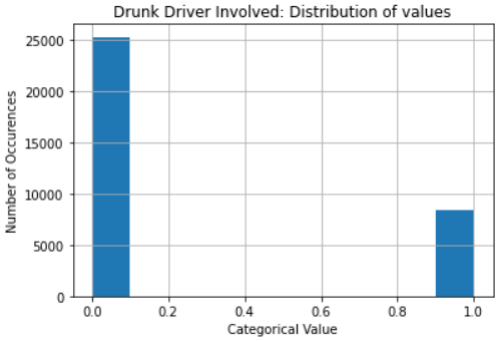

CF1: Distribution of values
Number of Occurences
Categorical Value


CF2: Distribution of values
Number of Occurences
Categorical Value


CF3: Distribution of values
Number of Occurences
Categorical Value


FATALS: Distribution of values
Number of Occurences
Categorical Value


MinutesOfDay: Distribution of values
Number of Occurences
Categorical Value


Drunk Driver Involved: Distribution of values
Number of Occurences
Categorical Value