Predicting Accident Severity in Indiana

Amanda Ezell

October 2020

# Introduction

1. **Background**

Everyday drivers are at risk of being in a traffic accident. Some are worse than others, and many would think that weather and road conditions could come into play in the resulting severity of a crash. Knowing conditions that could lead to more serious accidents could allow drivers to change trip plans to avoid a high potential for injury.

1. **Problem**

Several conditions could lead to a more severe accident causing injury. These could include light conditions, weather conditions, surface conditions, the type of vehicle, and the safety equipment in place. By analyzing previous accidents within Indiana, insight may be gained into the possible severity of future accidents in the state. With this knowledge, drivers can make adjustments to their route, driving style, or reassess the necessity of their trip.

1. **Interest**

Any driver planning a trip within Indiana could be interested in the findings to help avoid injury in an accident.

# Data Acquisition and Cleaning

1. **Data Source**

Data for all Indiana traffic accidents are available to the public by the State of Indiana. It is separated by year. I chose to use the traffic accident data from 2019.

Data can be found at the following link:

<https://hub.mph.in.gov/dataset/aries-crash-data-2007-2017/resource/7081439e-33c9-45f5-9c0b-86d8a8c49252?view_id=f56f9538-0d0b-4162-bd5a-09082e412f5f>

The data contains 625,702 rows and 107 columns. The data contains a large variety of columns surrounding each accident. Examples of columns included are: Gender, Age Group, Injury Status, Seatbelt use, Nature of Injury, Local of accident, Day of the week, Light conditions, Weather conditions, conditions, Surface conditions, Type of vehicle, Type of Road, etc. Not all columns have complete data. For each column that will be used in the prediction, the data will be cleaned to avoid skewing the prediction. The Injury Status Code column indicates the severity of the accident and has a range of 1-7. A status of 1 indicates a fatal accident, with the injury being less severe as it increases. Blanks indicate no injury, and those will be populated with an 8 as part of the data preparation.

1. **Data Cleaning & Feature Selection**

As indicated above, the Injury Status column is blank when there is no injury. Null data in the Injury Status column was updated to ‘NO INJURY’. In addition, the Injury Status Code column null values were updated to 8. Several of the potential target columns for prediction have associated numeric code columns. After populating the null values Injury Status Code and Description, a Pearson correlation was executed to narrow down Columns with potential. Evaluation of this correlation lead to reducing the columns in the dataframe to include the codes and descriptions for the following:

1. Injury Status
2. Safety Equipment Used
3. Light Conditions
4. Weather
5. Surface Type Condition
6. Road Type (ex. Multi-Lane, One-Way, etc.)
7. Vehicle Type

Null data in the resulting columns was then populated with an appropriate number for the code and ‘Unknown’ for the description. Once all null values were handled, another Pearson correlation was executed to see if there was any change. One Hot Encoding was used to convert the categorical data to binary prior to normalization of the data for Classification. Further analysis of the features was done during the exploratory analysis and predictive modeling phase.

# Methodology

1. **Exploratory Analysis & Predictive Modeling**

For the initial analysis, the aim was to predict the severity of the accident and not just if injury occurred. Could the models predict the various severities (i.e. Fatal, in-capacitating, etc) of accidents? As part of the analysis, Decision Trees and Logistic Regression were used. It was found that SVM and KNN were not efficient to process the large amount of data in the dataset.

1. Decision Tree

The Decision Tree model using features Safety Equipment Used, Light Conditions, Weather, Surface Type Condition, Road Type (ex. Multi-Lane, One-Way, etc.), and Vehicle Type provided accuracy of 93.67%. Checking the unique values of the prediction vs. the actual, it does appear to be a very good model. The tree was able to have a predictive path for each status with the exception of ‘Refused’ and ‘Unknown’, which is expected.

1. Logistic Regression

The Logistic Regression model using features Safety Equipment Used, Light Conditions, Weather, Surface Type Condition, Road Type (ex. Multi-Lane, One-Way, etc.), and Vehicle Type provided accuracy of 94%. Looking at the confusion matrix (Figure 1), we can see that no predictions were able to be made for Fatal or Possible injury, while the Decision Tree above could. The precision was very low for all possibilities with the exception of ‘NO INJURY’. See Figure 2.

Figure 1:

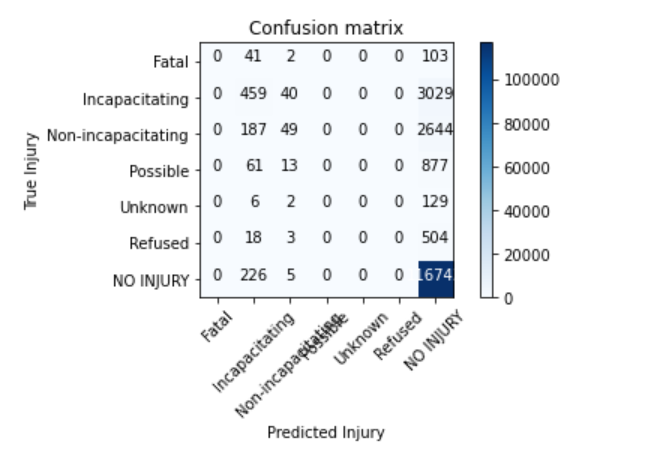
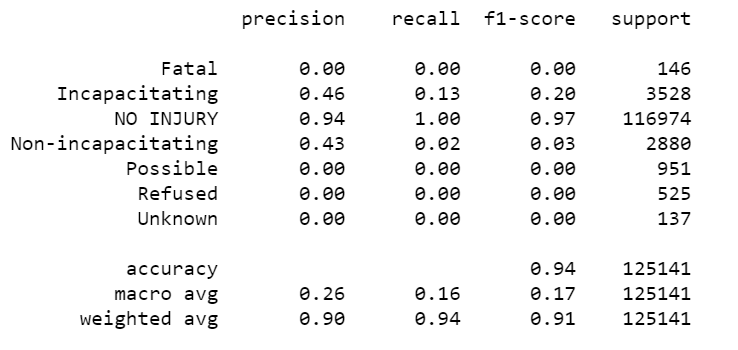


Figure 2:



Due to the lack of precision on the individual severities, I evaluated 2 additionla Logistic Regression models with fewer input features. As you can see in Figures 3-6, accuracy was maintained but precision was not improved.

Features: Safety Equipment Used, Light Conditions, Weather, Surface Type Condition, and Vehicle Type

Figure 3:

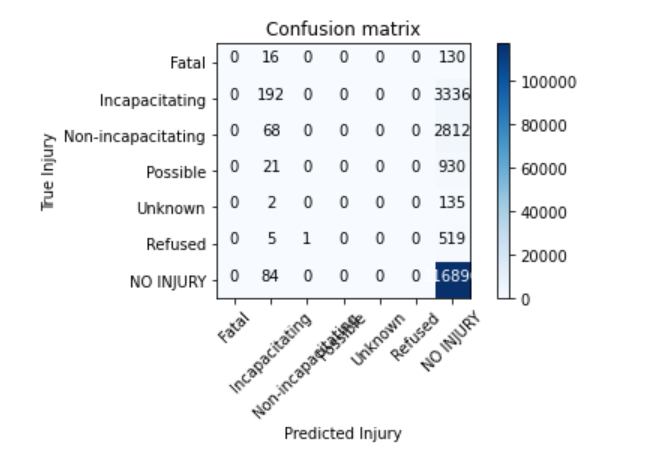
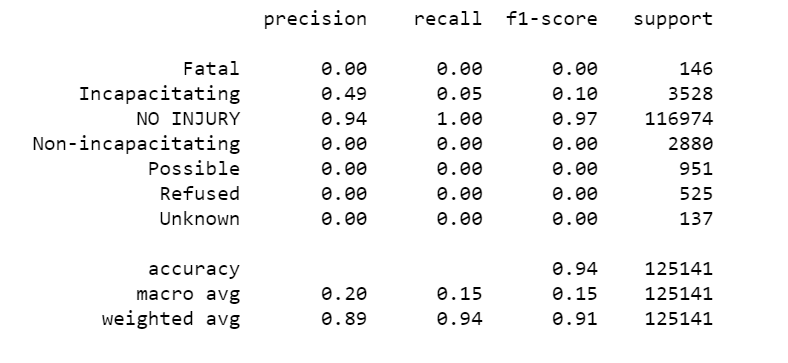


Figure 4:



Features: Safety Equipment Used, Light Conditions, Weather, and Vehicle Type

Figure 5:

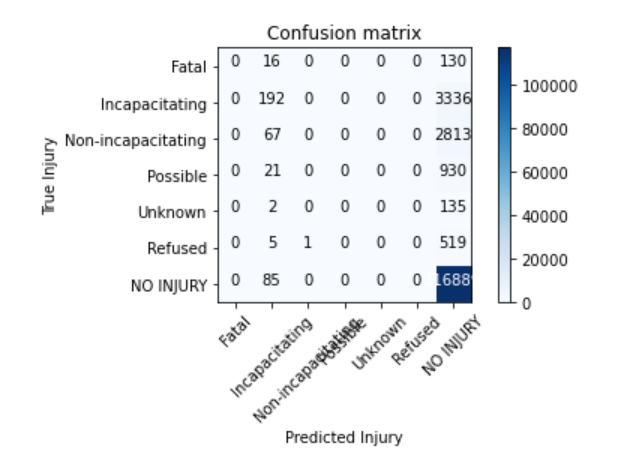
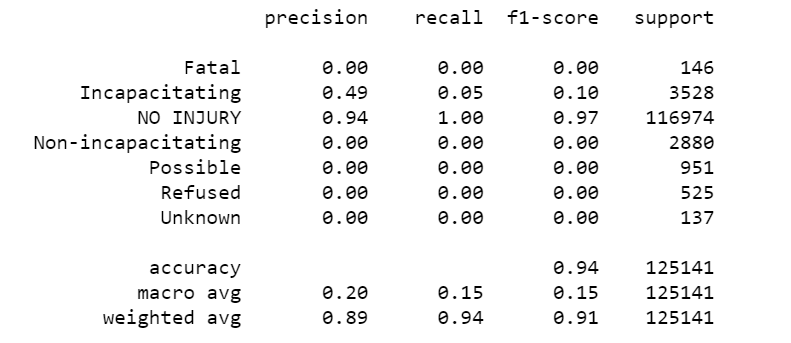


Figure 6:



1. **Exploratory Analysis & Predictive Modeling with binary outcome**

Due to the precision issues in the above Logistic Regression models, I pondered if there would be value in predicting ‘INJURY’ vs. ‘NO INJURY’ instead of the various levels of severity because knowing probability of injury may also lead to changes in behavior and be of value to some travelers.

To accomplish this, I created a new dataframe from the original dataset. Values in the Injury Status columns were then updated to ‘INJURY’ or ‘NO INJURY’ based on the original value. ‘Unknown’ and ‘Refused’ were mapped to ‘NO INJURY’ while all others were mapped to ‘INJURY’. Decision Tree and Logistic Regression was then performed.

1. Decision Tree

The same original set of features was picked for the predictive modeling. For clarity, features are Safety Equipment Used, Light Conditions, Weather, Surface Type Condition, Road Type (ex. Multi-Lane, One-Way, etc.), and Vehicle Type. The Decision Tree accuracy with the binary outcome possibility is 94.60%, which is slightly better than with the original with 7 possible outcomes.

1. Logistic Regression

The original set of features was also used for the Logistic Regression: Safety Equipment Used, Light Conditions, Weather, Surface Type Condition, Road Type (ex. Multi-Lane, One-Way, etc.), and Vehicle Type. We can see from Figure 7 and Figure 8 below, the accuracy remains at 94% and precision is improved.

Figure 7:

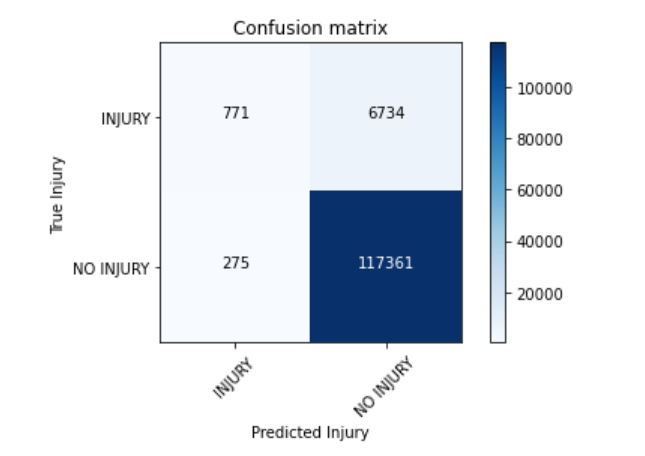
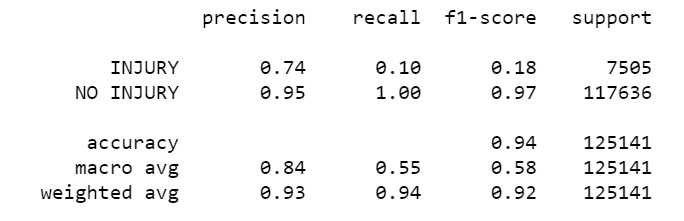


Figure 8:



1. **Model Evaluation**

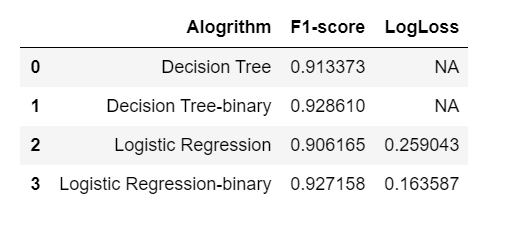
Both Decision Tree models and the Logistic Regression models with the original set of features were chosen to be evaluated. As Indiana has traffic data separated by year, I chose. I chose to use the traffic accident data from 2018.

Data can be found at the following link:

<https://hub.mph.in.gov/dataset/aries-crash-data-2007-2017/resource/cc90589c-72d8-4d92-a5fe-73254b555c73>

The data was cleaned and prepared like the previous data set. The model evaluation revealed that in both scenarios the Decision Trees performed slightly better. See Figure 9.

Figure 9:



# Results

We can see that the Decision Tree and Logistic Regression models do a very good job of predicting the severity of a traffic accident whether for the full result possibilities or in just determining Injury vs. No Injury.

# Discussion

It should be noted that overall a small percentage of accidents result in Injury in the State of Indiana. In the dataset from 2019, only 5.9% of all accidents resulted in injury. Similarly, in 2018 6.1% of all accidents results injury. Overall, a majority of accidents do not result in injury in the State of Indiana.

# Conclusion

Accident severity can be predicted with very good accuracy based on Safety Equipment Used, Light Conditions, Weather, Surface Type Condition, Road Type (ex. Multi-Lane, One-Way, etc.), and Vehicle Type. Travelers on Indiana roads can use either Decision Tree model based on their personal preference and predict the severity of the accident. They can utilize the first Decision Tree model and predict the level of injury if they were to be involved in an accident, or if that is too granular they can utilize the second Decision Tree model and predict the likelihood of injury if they were to be in an accident.