Predicting Car Accident Severity

By: Jonathan Klausner

1. Introduction

1.1 Background

Car accidents are very common on the road and happen every day across the country. In fact, Motor vehicle accidents are one of the leading causes of accidental death in the United States. Some accidents can be as small as lightly bumping into a parked car to six car pile-up on a major highway resulting in any amount of deaths and injuries. Insurance needs to be able to determine the severity of each accident to determine the amount of coverage they are offering. Outside forces such as weather, road and light conditions can help to predict how bad an accident could be as well as the frequency. Varying conditions can increase or decrease the likelihood of a more severe accident. Overall, public safety could benefit by this knowledge in order to encourage smarter decision making while on the road to avoid such accidents.

1.2 Business Problem

The desired result of this project is to be able to accurately predict the severity of a car accident based on a number of aspects, circumstances and features. The central aspect of the problem will revolve around light conditions and whether or not areas with limited to no light can cause more severe accidents.

1.3 Interest

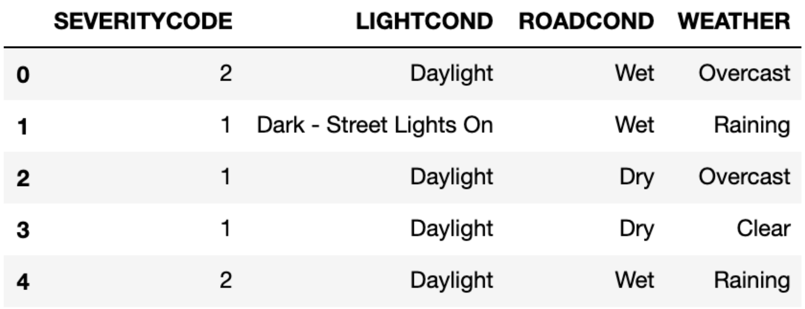
This problem could be of interest to anyone planning long car trips during questionable circumstances in order to avoid possible accidents. Truck drivers and other occupations that have people on the road often could use refer to this model and others to avoid future accidents. Anyone who drives a car would and should have an interest in accident severity statistics for safety reasons.

2. Data

The data being used in this project would be the shared data set from the Seattle Department of Transportation (SDOT). This data set contains information that is updated weekly and includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe: 2004 to Present. The data includes features such as counts of pedestrians, cars, etc. Useful features also include road and weather conditions, speed of the car, whether the drivers were under the influence and severity descriptions. The selected features will have to be tested for viability and connection to the selected dependent variable which will be SEVERITYCODE in this case. Variables such as INTKEY are not necessary to the actual detail of the accidents and also have a ton of missing inputs. Thus, variables like that will be omitted. The process will be broken down in the methodology section.

2.1 Variable Selection

Due to the problem's nature, the variable we would like to test for its correlation to SEVERITYCODE would be LIGHTCOND, ROADCOND, and WEATHER. These variables are the types of things that the average person would consider being factors to accidents and thus were selected to be tested for accuracy.A separate data frame was created with just these features for neatness and clear understanding for interpretation.



	SEVERITYCODE	LIGHTCOND	ROADCOND	WEATHER
0	2	Daylight	Wet	Overcast
1	1	Dark - Street Lights On	Wet	Raining
2	1	Daylight	Dry	Overcast
3	1	Daylight	Dry	Clear
4	2	Daylight	Wet	Raining

The data was cleaned by eliminating observations with Null values in order to convert the object datatypes to numerical int64. The dataset was then balanced by down sampling the larger subset of less severe accidents "SEVERITYCODE"==1 since there were more than twice the amount of less severe accidents than more severe accidents in the dataset.

3. Methodology

After the data was selected and cleaned, there was a K-Nearest Neighbor, Decision Tree and Logistic Regression model that was ran to test the accuracy of their predictions. The metadata for the shared dataset showed multiple values for the SEVERITYCODE feature but the actual data was binary between codes 1 and 2 thus the first test conducted was the logistic regression. After cleaning the data and eliminating null components the data was split between a training set and a testing set where 25% of the data was used to test and the remaining was used for training. The decision tree had an accuracy of 55.9% with its best test at a depth of 6. The K-Nearest Neighbors tried to find data points with similar attributes in order to predict and had an accuracy of about 55% when k is set to 20. This was created to test purely features that are out of the driver's control.

4. Results

The results of this problem show that attributes such as road conditions, light conditions and weather can all play a role in predicting the severity of a potential accident. These are all conditions that are environmental and thus out of control of the driver. The Jaccard similarity score of the logistic regression was about 0.55 showing a favoring accuracy that is above a 50% threshold. They do play a roll in predicting accidents when used at once. Separated and on their own they struggle to predict. Road conditions for example can help predict the likelihood of an accident occurring but not what severity the accident could be.

5. Discussion

The dataset was completely imbalanced due to the binary nature of the target variable. SEVERITYCODE had at least six different values but the actual data was classified by low severity accidents. Those with just property damage (1) and those with some or possible injury(2). A future study should balance out the data better and assign severity codes based off of more information such as fatalities, injury types, etc. which were not included in the data set. In terms of the evaluation, the different K's and depths were sifted through in order to find the best outcomes. The values used in logistic regression were F-1, Jaccard Index and Log Loss. For a better model, it should incorporate all features that are not redundant by repeating the same aspect or are just to label an incident. Things like speed, inattention, car and pedestrian account are all important but this is just using things like weather and road conditions.

6. Conclusion

In this study, I selected environmental features and tested to see how well they could be used in predicting the severity of an accident. I tested several machine learning models that could predict a severity code based off of the three selected features: light conditions (LIGHTCOND), road conditions (ROADCOND), and weather (WEATHER). This type of testing could be useful to those planning long road trips or for cities to be able to put out cautions or advisories to warn citizens about the potential for more severe accidents based on the current feature conditions.