Predicting the Severity of Road Collisions

Rachit Kumar

Saturday, October 3, 2020

1 Introduction

Car accidents cause property damage, are a huge disruption to traffic, and can cause trauma and even serious physical injury. Drivers want to avoid getting into accidents as much as possible, and perhaps even avoid driving when there may traffic disruptions due to an accident. We can use machine learning to predict the likelihood of a serious accident based on current weather, road, and lighting conditions. A system in the driver's car can then use this model by inputting the current conditions, and giving the driver a prediction of the likelihood of a severe accident. The driver can then use this information to decide whether or not to drive, and how carefully.

2 Data

The source of the data used in this analysis is the Seattle Police Department's Traffic Records database. It is a table with information about every recorded road collision in Seattle since 2004.

The data has a number of attributes including the geographic coordinates of the location of the collision, the address of the collision along with address type, the severity of the results of the collision, the type of collision, each of the number of people, pedestrians, cyclists, and motorists involved in the collision, the date and time of the collision, the weather conditions, the road conditions, the road lighting conditions, whether or not the driver was speeding, whether a parked car was involved, and a description of the collision according to the collision code.

3 Methodology

3.2 Data Exploration

The data was explored using the *pandas.describe* method.
Additionally, checks were made to see if there were any NaN values that needed to be discarded.

3.2 Machine-learning Model

The main target variable is to predict the severity code (1 or 2) of the collision with its probability of occuring. 1 means property damage only, no injury. 2 means the collision caused an injury. The probability will be expressed as a percentage.

The main independent variables are 'WEATHER', 'ROADCOND' and 'LIGHTCOND'. The machine learning model used will be logistic regression. This is the ideal algorithm because the severity code is binary and logistic regression can classify. A Decision Tree model will also be tested.

The independent variables will be coded as continuous variables. Here's how the WEATHER variable will be coded:

Clear	0
Partly Cloudy	1
Overcast	2
Fog/Smog/Smoke	3
Raining	4
Blowing Sand/Dirt	5
Severe Crosswind	6
Sleet/Hail/Freezing Rain	7
Snowing	8
Other	9
Unknown	9

The 9's will be discarded. This is because they represent incomplete or insufficient information. This same format (coding incomplete information as 9's) will also be employed for the remaining two independent variables. The reason this is done is to make the rest of the code simpler.

Here's how the ROADCOND variable will be coded:

Dry	0
Snow/Slush	1
Standing Water	2

Sand/Mud/Dirt	3
Oil	4
Wet	5
Ice	6
Other	9
Unknown	9

And lastly, here's how LIGHTCOND will be coded:

Daylight	0
Dawn	1
Dusk	2
Dark - Street Lights On	3
Dark - Street Lights Off	4
Dark - No Street Lights	5
Dark - Unknown Lighting	9
Other	9
Unknown	9

4 Results

The results show that Logistic Regression results in an accuracy of 0.6811393504941892. A Decision Tree machine learning model also resulted in an identical accuracy of 0.6811393504941892.

5 Discussion

The model was not very good at predicting the severity of the collision using weather, road conditions and lighting conditions. Both decision tree and logistic regression models resulted in low accuracies. However, the predictions are better than a blind guess (which would result in an accuracy of 50%. So it can be considered a partial success.

One of the possible reasons for the model not being very good is that the data do not include collisions that didn't occur. We do not have the traffic data for roads with given weather, road conditions and lighting conditions. For example, collisions on roads with snow were much more likely to be with no injury than all other types of collisions, including clear weather and wet weather. This is likely because in the snow, drivers are much more careful to drive very slowly. Thus, they can almost guarantee that there will not be a severe collision. However, it is very

easy to lose control of the vehicle and cause a minor collision. But we would likely see that during snowy days, there are more collisions per vehicle on the roads than on non-snowy days. The data completely misses this.

Therefore, one way to improve the model would be to gather data on the approximate number of vehicles on the roads under all given conditions. The collision data should then be normalised using that, to obtain a much better model with more accurate predictions.

6 Conclusion

The Logistic Regression model helps predict how severe a collision is going to be. Further data and testing is necessary to evaluate if this prediction is better than a driver's natural intuition.

A suggested improvement to the analysis is to include data for the total number of cars on the roads given the road conditions.