**Predicting the Severity of an accident**

Phani Sripada

September 20th, 2020

1. Introduction
   1. Background

There are about 40,000 people die every year in crashes on US roadways. The US traffic fatality rate is 12.4 deaths per 100,000 inhabitants. And additional 4.4 million are injured seriously enough to require medical attention. Road crashes are the leading cause of deaths in the U.S. for people aged between 1-54. When accident happens, emergency services such as police and medical rush to the spot and try to save the lives by providing them with required medical attention and in some cases the injured will be air lifted to the nearest hospital for further treatment. Depending on the location of the accident, it also results in huge traffic jams and the police and other public service personnel had to work round the clock to bring it back to normal.

* 1. Problem

There are several people get affected with each accident, starting from the people travelling in those vehicles involved in the collision, people whose property or vehicles got damaged or even people travelling on the same road to reach their destinations got delayed, public service agencies such as police, healthcare reach the spot etc.,

Given the seriousness and impact of each accident that sometimes leaves a significant impact on human lives, if there is a software that can predict the severity of an accident based on the weather, road conditions etc. and warn the users if they are going to get into an accident or traffic jam, so that they can drive carefully or choose alternate routes to reach their destinations safely.

* 1. Interest

This is very useful to the travelers using road as the mode of transport and by using this data they can take alternate routes when there is a high severity collision take paces. The government officials can add measures to make people alert while driving through some of the high severity accident zones.

1. Data Acquisition and Cleaning
   1. Data Sources

The data set that I have taken is an example data set that contains the details of collisions that occurred since 2014 that are recorded at Seattle Traffic management Division and Traffic records group.

* 1. Data Cleaning

There are total 37 various attributes available corresponding to each collision. However, there are a few features having empty data for most of the records. I analyzed the data closely and took appropriate actions against each one of them.

For those categorical columns missing values I grouped them into UKNOWN category and some cases where UKNOWN category did not exist, I ended up creating a new one. For ex: ADDRTYPE, INTKEY, LOCATION, JUNCTIONTYPE, INATTENTIONIND

There are a few columns which either have empty values of the existing data does not help in the analysis, I choose to drop them. For ex: EXCEPTRSNCODE, EXCEPTRSNDESC,.

* 1. Feature Selection

After cleaning the data, there are xxx records and xxx features left in the data set. Upon further examination, I noticed that there is some duplicate information such as a separate column to describe the ID present in the adjacent column. I choose the exclude such columns.

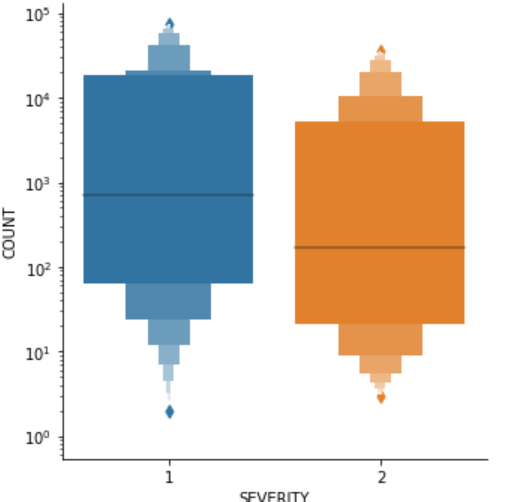
There are some features which represent only as an ID or key for the record and it does not provide any useful insight into the collision.

Finally, I picked the following features that are to be taken up for further data analysis.

1. Exploratory Data analysis
   1. Relationship between weather conditions and the Severity of the Collision

As we all know, weather plays an important role in our driving. It is not only that the extreme weather conditions can cause major accidents but also sometimes there is a clear sky, drivers tend overspeed and that can cause fatal accidents.

When I plotted a bar chart, it clearly gives an indication that the median is much below for Sev 2 accidents and lot of variance in the data distribution.

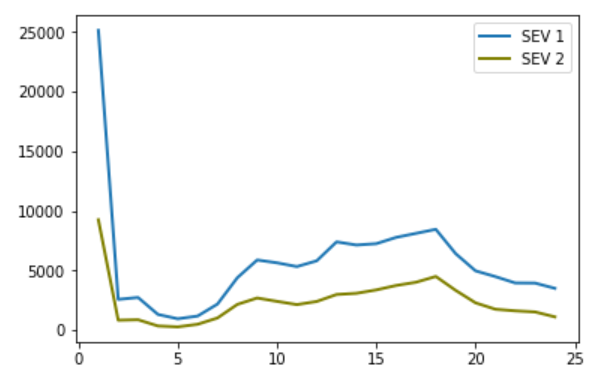


* 1. Relationship between hour of an accident and the Severity of the Collision

Since travelling in the night is riskier due to the lighting conditions and drivers tend to be drowsy leading to major accidents.

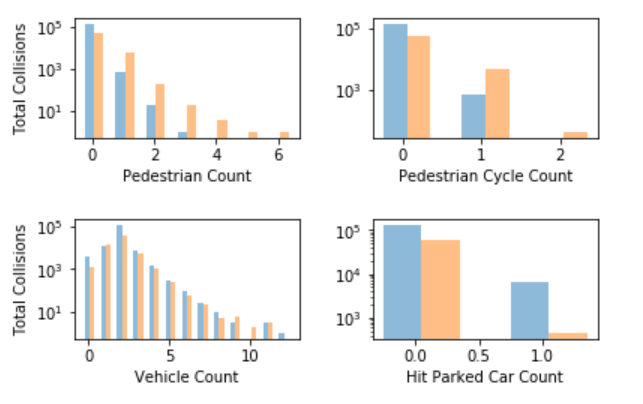
So, I extracted the hour of the collision from the INCTTIME provided in the data set and plotted a graph with hour as the x-asis and the number of collisions at each hour in a day.

Though the midnight collisions are very high in number for both Sev1 and Sev2, it is not clearly visible that whether a particular severity can be helpful in determining the severity.



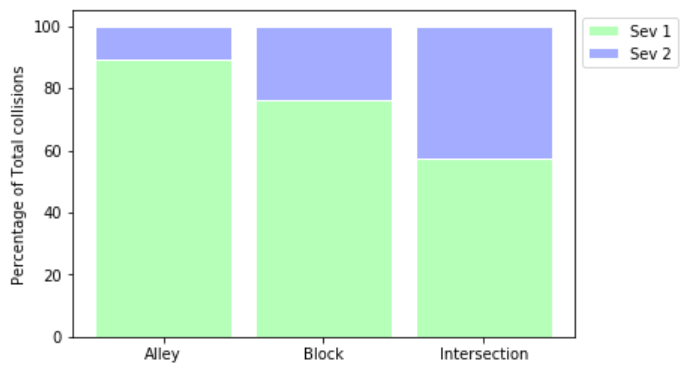
* 1. Relationship between pedestrian count and pedestrian cycle count and the Severity of the Collision

As per the visualization, whenever there is a pedestrian or pedestrian cycle involved in an accident, then that is most likely to end up being high severity collision.



* 1. Relation between ADDRTYPE and the Severity of the Collision

It is evident from the visualization that if the collision is of type Ally or Block then it is most likely to fall into Sev 1 collision, but it is difficult to predict when there is the collision is at an intersection.



* 1. Relationship between location of the collision and the Severity of the Collision

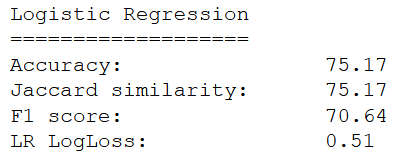
I felt it is worth examining the location coordinates that gives us an intuition on if the design of the road or any other distractions that are causing the driver to get into a major severity accident compared to other places.

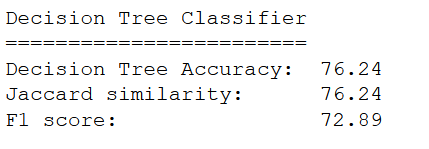
I plotted a Choropleth map to identify the locations by marking the top 10 accidents in each severity. I found that there is not much information that distinguishes locations based on their severities.

1. Predictive Modelling

As this is a classification problem, I choose two algorithms Logistic regression and Decision tree to predict the target and the accuracy scores corresponding to each model.

Performance of these models





1. Conclusions

There are certain features influencing the severity of an accident and also helping the model in predicting the severity of an accident. By taking advantage of this prediction, Users can avoid using those routes where the prediction of severity of an accident is high.

1. Future Directions

* It can be integrated with Google maps to alert drives travelling on that road to take conscious decisions.
* Government officials can further dig into the data to identify if there are certain roads becoming dangerous under circumstances such as bad weather, low lighting etc.
* First responders can use this data in getting the needed support to the victims as quickly as possible before reaching the site of the incident.